Estimating Uncertainty Models for Speech Source Localization in Real-World Environments

by

Kevin William Wilson

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Electrical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2006

Author .....
Department of Electrical Engineering and Computer Science
September 5, 2006

Certified by .....
Trevor Darrell
Associate Professor
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Estimating Uncertainty Models for Speech Source Localization in Real-World Environments

by

Kevin William Wilson

Submitted to the Department of Electrical Engineering and Computer Science
on September 5, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Electrical Engineering

Abstract

This thesis develops improved solutions to the problems of audio source localization and speech source separation in real reverberant environments. For source localization, it develops a new time- and frequency-dependent weighting function for the generalized cross-correlation framework for time delay estimation. This weighting function is derived from the speech spectrogram as the result of a transformation designed to optimally predict localization cue accuracy. By structuring the problem in this way, we take advantage of the nonstationarity of speech in a way that is similar to the psychoacoustics of the precedence effect. For source separation, we use the same weighting function as part of a simple probabilistic generative model of localization cues. We combine this localization cue model with a mixture model of speech log-spectra and use this combined model to do speech source separation. For both source localization and source separation, we show significantly performance improvements over existing techniques on both real and simulated data in a range of acoustic environments.

Thesis Supervisor: Trevor Darrell
Title: Associate Professor

Acknowledgments

I'd like to thank my advisor, Trevor Darrell, for starting me off on an interesting problem and then giving me the freedom to pursue my own approaches. I'd like to thank my committee members, Mike Brandstein, John Fisher, and Victor Zue, for sharing their perspectives and advice. Additional thanks to Mike for his help and for his input on early versions of this thesis.

I'd like to thank my friends at lab, including but not limited to Ali, Biswajit, David, Kate, Kinh, Louis-Philippe, Mario, Mike, Neal, Sybor, Tom, and Vibhav, for their support and for making MIT not only the center of my academic life but also the enjoyable center of my social life. Additional thanks to Biswajit, Mike, and Sybor for proofreading this document.

I'd like to thank my parents and my sister for supporting and encouraging me during my time at MIT as they have throughout my life.

Finally, I'd like to thank my girlfriend, Minnan, for her support and companionship and for being a great friend in every way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation makes contributions on the topics of audio source localization and simultaneous speech source separation in reverberant environments. Our primary claim is that by learning to predict audio localization cue reliability from time-frequency energy patterns, we can achieve more accurate source localization in reverberant environments. (A "localization cue" is a feature of an audio signal that depends upon the source location and can be used to estimate that location.) Secondarily, we will show that knowing cue reliability allows us to combine localization cues with monaural speech models to separate simultaneous speakers.

1.1 Motivation and goal

We rely heavily on our sense of hearing to understand the world around us. We organize our acoustic world by segmenting it into different auditory streams [15], where each stream typically represents sounds generated by a single entity, for example a person talking, a phone ringing, or a dog barking. Each stream at any given time will possess a number of perceptual attributes such as loudness, timbre, pitch (for harmonic sounds), and location. These perceptual attributes have strong relationships to physical properties of the sound sources, such as physical size, material properties, and position relative to the listener, and it is largely because of these relationships that hearing is so useful. For example, we can easily tell the the difference in sound be-

tween a small metal piccolo and a large wooden double-bass or between a Chihuahua and a Great Dane.

While both hearing and vision allow us to sense things at a distance, when an object of interest is visually occluded or when the object is out of our visual field of view, only our sense of hearing informs us about that object. These properties of hearing have helped to make spoken language the primary way that people communicate with each other, and for similar reasons, our environment has been designed to take advantage of our sense of hearing. Emergency vehicle sirens and honking car horns are examples of applications that make excellent use of our ability to immediately recognize and respond to a sound in spite of whatever else we may be focused on at the time [86].

Because of these unique properties of our sense of hearing, and because so much of the world around us is designed to take advantage of them, it is of great practical importance to better understand the sense of hearing and to find better engineering solutions for incorporating a sense of hearing into automated systems. The problem of endowing automated systems with a sense of hearing is very broad, however, and some aspects of the problem, such as timbre processing, are difficult to quantify and evaluate even though they are an important part of our auditory experience.

We choose to focus on what is usually a well-defined problem with a well-defined answer, the problem of audio source localization. In most situations, the physical position of the sound source is the natural correct answer to the question of where a sound is located. However, even the source localization problem does not always have an obvious solution. For example, if a listener is in one room and a sound source is in another room down the corridor, should the listener localize the sound to the corridor or to the other room? How much worse is one answer than the other? This thesis avoids these questions by focusing on the case where there is an unobstructed direct path between the source and the listener. Even in this case, we will see that there are interesting problems to solve. (We mention this subtlety primarily to point out that when dealing with complicated natural environments, problems are seldom as simple and well defined as we would like.)

In addition to being reasonably well-defined, source localization is also very useful. For example, successfully localizing a honking car horn can enable a pedestrian to take action to avoid being run over, so in this case, the location information itself is immediately useful. In other cases, such as audio stream segregation, source localization cues are useful not because we fundamentally care about the locations of the sources, but simply because sources in different locations will provide different localization cues. These different localization cues, in combination with other sound attributes, allow people to focus on one sound source among many even in noisy environments [19].

For all of these reasons, the goals of this thesis are to improve source localization accuracy and to demonstrate the utility of this improved localization for audio stream segregation.

1.2 The challenge

As we will discuss in more detail in Chapter 2, source localization of a single source with known stationary signal statistics in anechoic conditions is a problem with a well-understood signal processing solution.

These idealized conditions are violated in many practical scenarios, however, and it is in these violations that many of the interesting research questions lie. The two main violations upon which this thesis will focus are:

1. Speech is a nonstationary signal with complicated temporal dynamics. Existing source localization techniques have known error bounds and can be shown to be optimal for a stationary signal with a known spectrum, but these bounds are of limited use with speech signals whose spectra can change drastically and abruptly, for example from a vowel "a" to the stop consonant "b" in the word "abruptly."

2. Typical environments are not anechoic. In fact, once a receiver is more than a few meters from a source, more of the received energy will usually come from

reflections than will come from the direct path component. Reflections of a sound may also continue to arrive for hundreds of milliseconds after the direct path component, which means that energy from a few consecutive phonemes will be mixed together in the reverberated audio, further complicating the modeling of speech spectra.

As we will discuss further in Chapter 2, empirical results [18] do indeed show that reverberation significantly degrades localization performance. Other work [38] suggests that this performance is approximately optimal for stationary signals, and that localization in reverberant environments is a fundamentally more difficult problem than localization in anechoic environments.

Still, there is room for improvement, since psychoacoustic evidence [8] and everyday experience show that people can localize sounds reasonably well in reverberant environments.[1] This suggests that there is additional structure in the source localization problem that we can exploit, and I will argue that this additional structure lies in the interaction between the nonstationary nature of speech and the acoustics of reverberant environments.

1.3 The Basic Idea

The basic premise of this thesis is that low-level general-purpose localization cues (which for our purposes will be phase differences in time-frequency signal representations) do a reasonable job of summarizing the localization information in small time-frequency regions, but that because we do not know how reliable the information in each time-frequency region is, it is difficult to combine these cues into a good overall location estimate. Combining cues is critical because individual cues may be noisy or fundamentally ambiguous.

[1] It is difficult to find directly comparable data for human and machine localization performance in reverberant environments. The most directly comparable results of which we are aware are that [38] shows a total root-mean-square (RMS) time delay error of over 40 microseconds for their technique on a 50 millisecond window of broadband noise in a reverberation time of 0.4 seconds, while human performance localizing a 50 millisecond tone burst is below 35 microseconds RMS error at reverberation times as high as 4 seconds [40].

What we need, then, and what this thesis develops, is a technique for predicting the reliability of these low-level cues. In particular, we set out to find some observable features of the signal that correlate with cue reliability. We then learn the relationship between these features and cue reliability and exploit this relationship to better combine low-level cues across time and frequency.

Since humans localize sounds well even in difficult environments, we look to the psychoacoustics literature for hints about what sorts of features might be useful for this task. There we find the "precedence effect," in which humans rely more heavily on localization cues from sound onsets and suppress cues from steady-state sounds [59] in order to emphasize parts of the signal that are less corrupted by reverberation. This effect is potentially useful and is lacking from the current signal processing approach, so we will seek to formalize this effect and incorporate it into an appropriate signal processing framework.

1.4 Contributions

This thesis contributes improved techniques for audio source localization and speech source separation in reverberant environments.

For source localization, this thesis makes two contributions. First, we use a training corpus of reverberated speech and associated localization cues to learn a mapping from the speech signal to a measure of localization uncertainty, and we relate this procedure to maximum likelihood time delay estimation. Second, we make a connection between the mappings learned by our system and the precedence effect.

For source separation, we combine our model of localization cue reliability with a simple speaker-independent speech model to separate simultaneous speech sources.

For both localization and separation, we demonstrate significant performance improvements over existing techniques across a wide range of acoustic environments.

Portions of this work have been published in [90–92].

1.5 Structure of the Dissertation

Chapter 2 reviews related background on audio source localization from both signal processing and psychoacoustic perspectives. Chapter 3 describes our source localization technique, which is the primary intellectual contribution of this work. Chapter 4 describes the experimental evaluation of our localization technique. Chapter 5 describes how localization cues can be used to separate simultaneous speech sources along with experimental results for this task. Chapter 6 summarizes this work and suggests future directions.

Chapter 2

Background

This thesis work is motivated by psychoacoustic findings and statistical signal processing theory. This chapter reviews the basics of these subjects as they relate to our work. Specifically, this chapter addresses the following questions:

1. In theory, how should one optimally estimate the location of a sound source? Under what conditions can we guarantee optimality?
2. How does reverberation affect source localization?
3. How do humans localize sounds, and how do they deal with the effects of reverberation?

The essence of this problem, in both psychoacoustics and signal processing, is captured by the two-sensor case. Binaural cues are the main source of localization cues in all land mammals that have been studied [33] and in nearly all other known animals. Also, stereo audio and stereo recording equipment are widely available, making this an important practical case. For these reasons, we focus on the two-microphone (or two-ear) case in the remainder of this dissertation.

2.1 Signal processing background

Two (or more) sensors in known relative positions are referred to as a sensor array. There is an extensive literature describing all of the exciting things that can be done

with arrays (see [51] for a good general array processing textbook and [53] for a general review article). However, our focus is on source localization and specifically source localization in reverberant environments, and we will focus our overview to that topic.

The goal of the array processing approach to source localization is to determine location based on differences in the received signals at different sensors. To explain the theory, we begin with the simplest possible model, two sensors in free-field nondispersive conditions (conditions in which waves propagate without interacting with obstacles and in which waves of all frequencies travel with the same velocity), and we attempt to localize a point source. Under these conditions, our source signal travels spherically out from the source location at a constant velocity and undergoes $1/r$ amplitude attenuation. Figure 2-1 shows an example with two sensors on the y-axis centered about the origin. In this case, we have

$$x_i(t) = \frac{1}{r_i} s(t - \frac{r_i}{v}) \tag{2.1}$$

where $x_i(t)$ is the observed signal at sensor i, $i \in \{1, 2\}$, r_i is the distance from the source to sensor i, $s(t)$ is the source signal, and v is the signal propagation velocity. The received signals differ by a shift and scaling depending on their distance from the source.

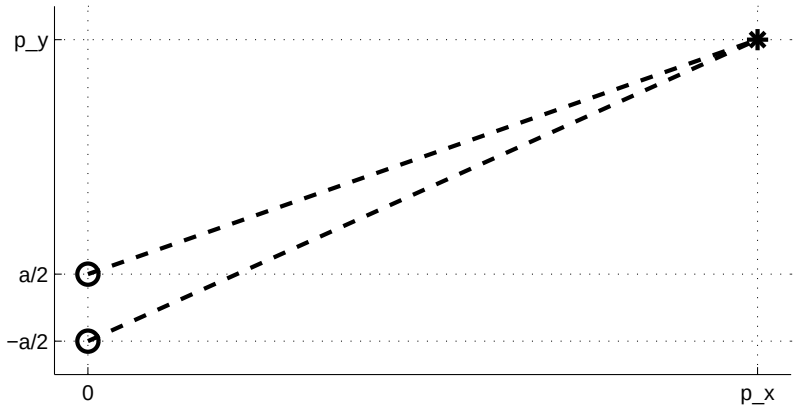
A useful special case is that of far-field sources, where $r_i \gg a$, with a denoting the separation between the microphones. In this case, because $|r_1 - r_2| \leq a$ (by the triangle inequality), $1/r_1 \approx 1/r_2$ and the amplitude differences between the two channels becomes negligible. As depicted in Figure 2-1(b), in this case the directions of signal propagation from the source to each microphone are nearly parallel, and the path length difference is approximately $a \sin \phi$, where ϕ is the direction of arrival. Path length difference is directly proportional to time delay with constant of proportionality $1/v$, so the angle of arrival is

$$\phi = \arcsin(\frac{vD}{a}) \tag{2.2}$$

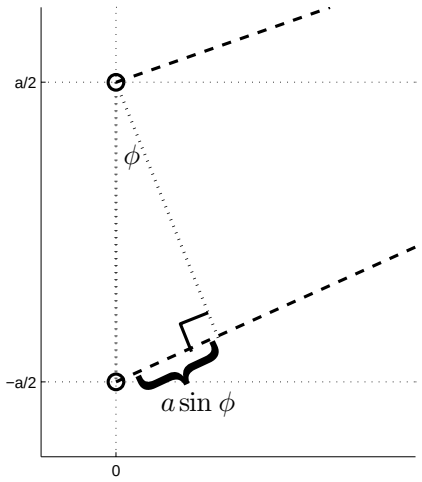
$$D = \frac{r_2 - r_1}{v} \tag{2.3}$$

where D is the time delay between the two microphones. Therefore we see that there is a simple relationship between time delay and direction of arrival. Because the time delay depends on only the direction, and because the amplitude differences between the channels are negligible for far-field sources, there is no way of determining the distance to a far field source; the direction of arrival is all we can know. Because of the close relationship between time delay and direction of arrival, and because direction of arrival is the only information about the location that we can recover in the 2-microphone far-field free-field case, we will speak somewhat interchangeably of source localization, direction of arrival estimation, and time delay or time-delay-of-arrival (TDOA) estimation throughout this dissertation.

Time delay is what we need to localize (determine the direction of) a source, so now we must figure out how to estimate it from our received signals. From Equation 2.1 we can derive that $x_1(t) = \frac{r_1}{r_2}x_2(t - D)$. This means that $R_{x_1x_2}(\tau) = \frac{r_1}{r_2}R_{x_1x_1}(\tau - D)$, where $R_{x_ix_j}(\tau)$ is the cross-correlation function of $x_i(t)$ and $x_j(t)$ at time-lag τ. It is a general property of the autocorrelation (the cross-correlation of a signal with itself) that $R_{xx}(\tau) \leq R_{xx}(0)$ [70, p. 818], or in other words the autocorrelation reaches a (possibly non-unique) global maximum at zero lag. Because of this, $R_{x_1x_2}(\tau)$ will reach a maximum at D, so we can determine the time delay between two signals by locating the peak in their cross-correlation. (Note: If $s(t)$ is periodic with period T, $R_{x_1x_2}(\tau)$ will also be periodic with period T and will therefore have no unique global maximum. One of its infinitely many peaks that reach the global maximum value will occur at D, however, and it may be possible to uniquely identify the peak at lag D based on constraints imposed by the array geometry, since the maximum possible delay for a propagating signal is a/v.)



(a)



(b)

Figure 2-1: Geometric view of the relationship between time delay estimation and source localization in two dimensions. (a) shows a distant source as an asterisk and two microphones as circles. (b) is a close-up view of the sensors that shows the relationship between the path length difference and the angle of arrival.

2.1.1 Generalized cross-correlation

In the noise-free case, estimating the time delay requires finding the maximum in the cross-correlation function. When ambient noise is present it would seem advantageous to emphasize parts of the signal with high signal-to-noise ratio (SNR) and suppress parts with low SNR. Knapp and Carter [52] analyzed this problem, and formalized this intuition. They assume an additive noise model, i.e.

$$x_1(t) = s(t) + n_1(t) \tag{2.4}$$

$$x_2(t) = s(t - D) + n_2(t) \tag{2.5}$$

where $n_i(t)$ is a noise term and for simplicity we omit the distance-dependent amplitude coefficient. Knapp and Carter assume that $s(t)$ and $n_i(t)$ are zero-mean stationary Gaussian random processes and that the noise is uncorrelated with the source signal and across channels. Their analysis is done in the frequency domain on finite-length observations with

$$X_i(f) = \frac{1}{T} \int_{-T/2}^{T/2} x_i(t) e^{-j2\pi ft} dt \tag{2.6}$$

$$G_{x_i x_j}(f) = \frac{1}{T} \int_{-\infty}^{\infty} R_{x_i x_j}(\tau) e^{-j2\pi f\tau} d\tau \tag{2.7}$$

where T is the observation window length, $X_i(f)$ is the Fourier transform of $x_i(t)$, and $G_{x_i x_j}(f)$ is the cross spectral density (the Fourier transform of $R_{x_i x_j}(\tau)$). They then define the generalized cross-correlation (GCC) function as

$$R_{y_1 y_2}(\tau) = \int_{-\infty}^{\infty} \Psi(f) G_{x_1 x_2}(f) e^{j2\pi f\tau} df \tag{2.8}$$

where $\Psi(f)$ is a frequency-dependent weighting function. Note that we have changed cross-correlation subscripts from x to y to indicate that this is no longer the cross correlation of the original $x_i(t)$. Knapp and Carter go on to show that for the

case of long observation window length, the maximum likelihood (ML) time delay estimator can be expressed as

$$\hat{D}_{ML_{GCC}} = \arg\max_t \int_{-\infty}^{\infty} \Psi_{ML}(f) G_{x_1x_2}(f) e^{j2\pi ft} df \tag{2.9}$$

$$\Psi_{ML}(f) \triangleq \frac{1}{|G_{x_1x_2}|} \cdot \frac{|\gamma_{x_1x_2}(f)|^2}{[1-|\gamma_{x_1x_2}(f)|^2]} \tag{2.10}$$

$$\gamma_{x_1x_2} \triangleq \frac{G_{x_1x_2}(f)}{\sqrt{G_{x_1x_1}(f)G_{x_2x_2}(f)}} \tag{2.11}$$

where $\hat{D}_{ML_{GCC}}$ will denote the ML estimate of the delay. $\Psi_{ML}(f)$ is the weighting function used to compute this estimate expressed in terms of $\gamma_{x_1x_2}(f)$, the interchannel coherence function, which is a complex-valued generalization of the correlation coefficient. Equation 2.10 is the most common way of expressing $\Psi_{ML}(f)$ in the literature, but based on our additive signal-plus-noise model (Equation 2.5), it can be re-expressed as

$$\Psi_{ML}(f) = A(f) \cdot B(f) \tag{2.12}$$

$$A(f) \triangleq \frac{1}{|G_{x_1x_2}(f)|} \tag{2.13}$$

$$B(f) \triangleq \frac{|\gamma_{x_1x_2}(f)|^2}{[1-|\gamma_{x_1x_2}(f)|^2]} \tag{2.14}$$

$$= \frac{G_{ss}^2(f)}{[G_{ss}(f)+G_{n_1n_1}(f)][G_{ss}(f)+G_{n_2n_2}(f)]-G_{ss}^2(f)} \tag{2.15}$$

$$= \frac{G_{ss}^2(f)}{G_{ss}(f)G_{n_1n_1}(f)+G_{ss}(f)G_{n_2n_2}(f)+G_{n_1n_1}(f)G_{n_2n_2}(f)} \tag{2.16}$$

First note that the $A(f)$ term is whitening the cross power spectrum of $X_i(f)$ since $G_{x_1x_2}(f)/|G_{x_1x_2}(f)| = 1$ for all f.

Next, if we assume $G_{n_1n_1}(f) = G_{n_2n_2}(f) = G_{nn}(f)$ and $G_{nn}(f) \gg G_{ss}(f)$, we have

$$B(f) \approx \frac{G_{ss}^2(f)}{G_{nn}^2(f)} \tag{2.17}$$

$G_{ss}(f)/G_{nn}(f)$ is the SNR, so we see that for low SNR (when $G_{ss}(f) \ll G_{nn}(f)$), the ML weighting is approximately proportional to the squared SNR. This brings us back to our stated intuition that we should emphasize frequencies with high SNR.

Next, let us consider how $\Psi_{ML}(f)$ relates to the variance of the complex phase of our cross spectrum, $\text{var}[\angle\hat{G}_{x_1x_2}(f)]$. When we estimate delay, we only have a finite-length observation from which to estimate the cross spectrum, and we define this estimate to be

$$\begin{aligned}
\hat{G}_{x_1x_2}(f) &\triangleq X_1(f)X_2^*(f) && (2.18)\\
&= [S(f)+N_1(f)][S(f)e^{-j\theta(f)}+N_2(f)]^* && (2.19)\\
&= \hat{G}_{ss}(f)e^{j\theta(f)} + \hat{G}_{sn_2}(f)e^{j\theta(f)} + \hat{G}_{n_1s}(f) + \hat{G}_{n_1n_2}(f) && (2.20)\\
&= \hat{G}_{ss}(f)e^{j\theta(f)} + V(f) && (2.21)\\
V(f) &\triangleq \hat{G}_{sn_2}e^{j\theta(f)} + \hat{G}_{n_1s} + \hat{G}_{n_1n_2} && (2.22)\\
\theta(f) &\triangleq 2\pi f D && (2.23)
\end{aligned}$$

Equation 2.20 expresses the estimated interchannel cross spectrum as a sum of terms dependent on the target signal and noise. What we care about is the relative phase of the target signal components, which is only encoded in the first additive term of Equation 2.21. We group the remaining noise-contaminated terms into $V(f)$. (The phase difference, $\theta(f)$, appears in the first term of $V(f)$, but $\hat{G}_{sn_2}(f)$ is itself a complex vector with random phase, so $\hat{G}_{sn_2}e^{j\theta(f)}$ as a whole will have random phase.)

Assuming the noise is small, we can visualize this as in Figure 2-2. Here we show the first term in Equation 2.21 as a complex vector with length $|G_{ss}(f)|$ and angle $\theta(f)$. The three terms in $V(f)$ are all uncorrelated and have uniformly distributed random phase, so we can combine them into a single Gaussian noise term represented by the shorter arrow and small, dotted circle. The signal and noise terms will have expected magnitude

$$E[\hat{G}_{ss}(f)] = G_{ss}(f) \quad (2.24)$$

$$E[|V(f)|] = \sqrt{G_{ss}(f)G_{n_2n_2} + G_{n_1n_1}(f)G_{ss}(f) + G_{n_1n_1}(f)G_{n_2n_2}(f)} \quad (2.25)$$

For small error ($E[\hat{G}_{ss}(f)] \gg E[|V(f)|]$), the expected phase error magnitude will be

$$E[|\theta(f) - \hat{\theta}(f)|] \approx \frac{\sqrt{G_{ss}(f)G_{n_2n_2} + G_{n_1n_1}(f)G_{ss}(f) + G_{n_1n_1}(f)G_{n_2n_2}(f)}}{G_{ss}(f)} \quad (2.26)$$

which means the phase error variance will be

$$E[|\theta(f) - \hat{\theta}(f)|^2] \approx \frac{G_{ss}(f)G_{n_2n_2} + G_{n_1n_1}(f)G_{ss}(f) + G_{n_1n_1}(f)G_{n_2n_2}(f)}{LG_{ss}^2(f)} \quad (2.27)$$

where L is a constant of proportionality. The left hand side of Equation 2.27 is just $B^{-1}(f)$ from Equation 2.15. Thus we see that ML weighting is approximately equivalent to whitening the cross spectrum (the $A(f)$ term) and then weighting by the inverse phase variance (the $B(f)$ term). We will use this fact later. (The outline of a more rigorous but consequently more opaque derivation of this fact can be found in [50, p. 379].)

2.1.2 Delay estimation by regression on phase differences

We have never found Knapp and Carter's derivation particularly easy to visualize, so we will now briefly outline Piersol's work on a different time delay estimator that has the same asymptotic performance as the ML GCC weighting but a much more intuitive derivation [74]. Piersol's basic idea is that for two signals with relative delay D, their phase differences $\theta(f)$ as a function of frequency should be a line through the origin with a slope that depends on D:

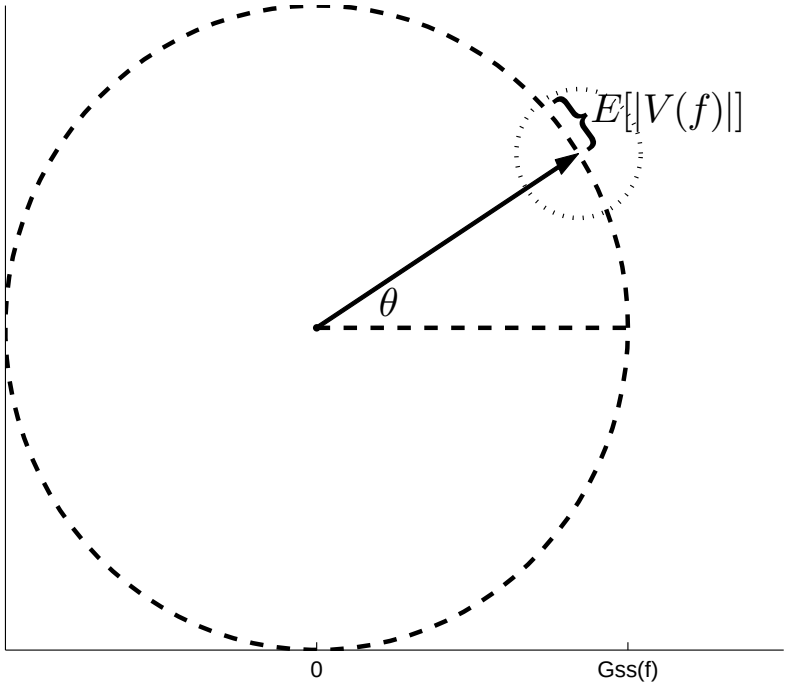


Figure 2-2: Graphical depiction of target signal phase difference and noise. The arrow from the origin represents the target-signal dependent component of the cross spectrum estimate. The small dotted circle shows the expected magnitude of the error.

$$\theta(f) = 2\pi f D \tag{2.28}$$

To find a delay, we compute the cross-spectrum at a discrete set of frequencies and find the slope of the weighted-least-squares best fit line through the cross spectrum phases as a function of frequency. (We also constrain the line to pass through the origin since real signals will always have zero phase at $f = 0$.) If we assume additive noise on the phases, then this is just a simple linear regression problem, whose general form and solution are

$$\begin{aligned}
\mathbf{\Theta} &\triangleq [\hat{\theta}(f_1) \quad \hat{\theta}(f_2) \quad \cdots \quad \hat{\theta}(f_N)]^{\mathsf{T}} && (2.29)\\
\mathbf{f} &\triangleq [f_1 \quad f_2 \quad \cdots \quad f_N]^{\mathsf{T}} && (2.30)\\
\mathbf{w} &\sim \mathcal{N}(0, \Lambda_\theta) && (2.31)\\
\mathbf{\Theta} &= 2\pi D\mathbf{f} + \mathbf{w} && (2.32)\\
\hat{D}_{ML_{Piersol}} &= \frac{1}{2\pi}(\mathbf{f}^{\mathsf{T}}{\mathbf{\Lambda}_\theta}^{-1}\mathbf{f})^{-1}\mathbf{f}^{\mathsf{T}}{\mathbf{\Lambda}_\theta}^{-1}\mathbf{\Theta} && (2.33)
\end{aligned}$$

where $\mathbf{w}$ is a phase observation noise process. (Because phase is periodic, this is only a simple linear regression problem if the phase measurements can be accurately unwrapped.) For appropriately chosen f_k, its covariance matrix, $\mathbf{\Lambda}_\theta$, will be diagonal, so we can rewrite this as

$$\hat{D}_{ML_{Piersol}} = \frac{1}{2\pi} \cdot \frac{\sum_{k=1}^{N} \frac{f_k \hat{\theta}(f_k)}{\mathrm{var}[\theta(f_k)]}}{\sum_{k=1}^{N} \frac{f_k^2}{\mathrm{var}[\theta(f_k)]}} \tag{2.34}$$

The summation in the denominator of Equation 2.34 is just a normalizing constant. In the numerator, we are combining phase estimates from different frequencies while weighting by their phase variances, just as we did in the GCC time delay estimator.

We have described Piersol's technique primarily to provide an alternative view of why phase error variance is important for time-delay estimation. As shown in [74],

both GCC and Piersol's time delay estimator have the same asymptotic performance. GCC is more commonly used in practice, and it avoids the need to explicitly deal with phase wrapping, so for our experiments we will focus on GCC-based methods.

2.2 Practicalities

In the previous section, we showed how to optimally estimate time delay of a stationary signal in uncorrelated noise with long observation windows where we also assumed that we knew the statistics of the signal and the noise. However, in many situations of practical interest, all of these assumptions are incorrect. This section explains how these assumptions are violated and how previous work has dealt with these violations in practice.

Observation window length is often limited by practical considerations. When the source is in motion, we need a window short enough that it allows little noticeable source movement; otherwise cross-correlation peaks corresponding to different source locations will be superposed, broadening the observed peak and reducing localization resolution. The nonstationary nature of many signals of interest, such as speech, is another reason to favor shorter windows. Speech is typically treated as quasistationary for time scales of a few tens of milliseconds, although even at those window lengths, some nonstationarity is apparent [85].

A bigger problem in practice, however, is that we do not know the source and noise signals' statistics. This difficulty is compounded by the fact that speech is nonstationary, so even making reasonable estimates of its statistics is difficult. In some cases, such as when the interfering noise also consists of speech, the noise process is also nonstationary, further complicating any attempt to estimate signal or noise statistics.

That leaves our "uncorrelated noise" assumption, which is violated in any reverberant environment. In such environments, delayed and possibly filtered copies of the original source signal arrive at the sensors from different apparent directions. For the purpose of localizing sounds, anything but the sound arriving via the direct path from

the source to the receivers should be considered "noise," so these reflections are noise, and because they are filtered versions of the target signal, they are clearly correlated with the source.

The good news is that even after violating all of these assumptions, GCC techniques can still work reasonably well in practice. In particular, a GCC weighting known as the phase transform (PHAT), has been found to work reasonably well in reverberant environments [38]. We will describe the phase transform shortly, but first let us discuss some of the properties of reverberation.

2.2.1 Reverberation

People's natural environment is not anechoic. Instead, sounds bounce off of or diffract around objects in the environment, so sound from a given source typically follows many distinct paths through the environment before reaching the listener.

A simplified but still very useful model is to imagine the source and listener in a rectangular room and to think of each of the walls as a "sound mirror" as depicted in Figure 2-3. This is the "image method" [3] of simulating reverberation, and it captures most of its important features. In this situation, the receiver receives a copy of the target signal from the physical target and from each of the virtual "image sources" that result from reflecting the true source location about one or more walls or virtual walls. The virtual sources are equivalent to the source reflecting off the wall. First order reflections are modeled by the boxes immediately adjacent to the physical room in Figure 2-3. These walls and virtual walls (depicted as dashed lines in Figure 2-3) absorb some fraction of the sound's energy each time it is reflected, so the virtual sources will be attenuated based on how far they have to travel along their virtual path and how many sound-absorbing virtual walls they have been reflected about.

Figure 2-4(a) shows an example impulse response generated by the image method. One feature of this impulse response is that in the first hundreth of a second, we see a number of well-separated discrete impulses, which represent discrete first-order reflections off of walls. When estimating time delays, these early reflections will

appear to come from the direction of the corresponding image source and will generate additional peaks in the cross-correlation function. This is one way that reverberation can cause time delay estimation errors.

Later in the tail of the impulse response, the image sources are more attenuated and more numerous. These late reflections may be well-approximated by an exponentially-decaying noise process. (The exponentially decaying "tail" is most obvious in the log-magnitude domain, as shown in Figure 2-4(b).) When estimating time delays, this tail is unlikely to cause distinct peaks in the cross-correlation function, but it will serve to increase the overall effective noise level. This exponential behavior exists because for longer delays, the image sources will have on average been reflected off of more virtual walls, and each virtual wall absorbs a constant fraction of the signal's energy. The dashed line in Figure 2-4(b) shows the best-fit slope, and this slope gives us a convenient way to characterize reverberation. We will use the common definition of the "reverberation time" of a room as the amount of time it takes for the reverberant energy to decay by 60 dB, and will often refer to this as the RT_{60}.

2.2.2 The phase transform

It has been experimentally observed that a particular GCC weighting function, the phase transform (PHAT), works reasonably well in reverberant environments [23]. The PHAT weighting is defined as

$$\Psi_{PHAT}(f) = \frac{1}{|\hat{G}_{x_1x_2}|} \tag{2.35}$$

First note that the PHAT weighting depends only on the observed signal statistics. Unlike the ML weighting, it does not depend on the (typically unknown) noise and target signal statistics. This makes it implementable in practice, unlike the ideal ML weighting. Next note that $\Psi_{PHAT}(f)$ is equal to $A(f)$ from Equation 2.13; it is just whitening the cross power spectrum. It would be the ML weighting if we had $B(f) = 1$. $B(f)$ is a constant, and since it is the term that depends on the SNR, the

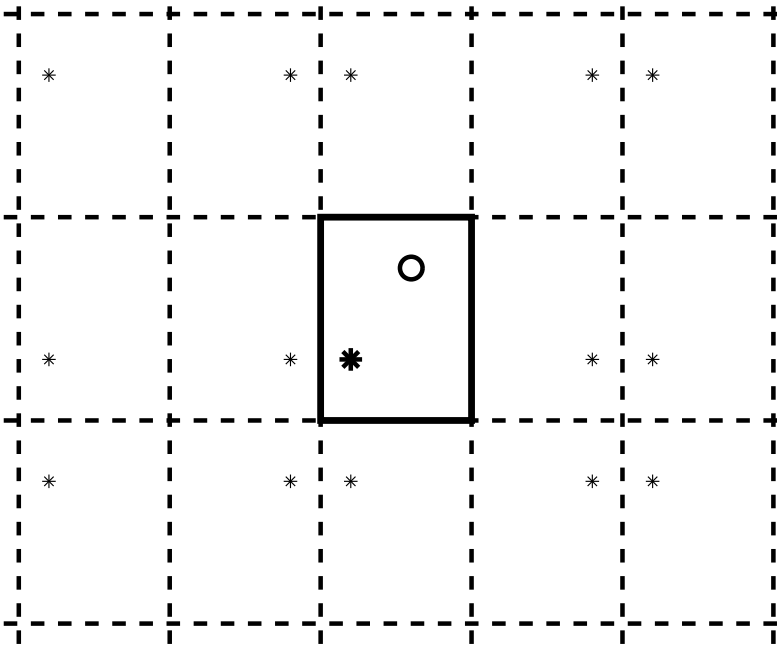
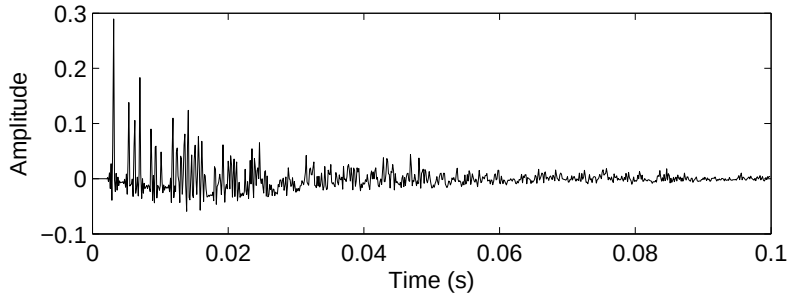
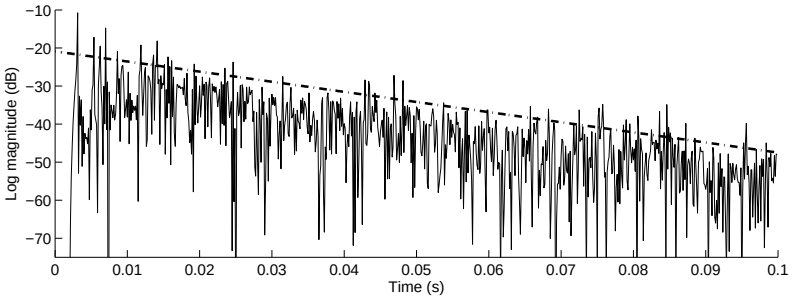


Figure 2-3: Two-dimensional example of the image method of simulating reverberation. The physical room is in the center and contains the physical source and receiver, denoted by the large asterisk and the large circle, respectively. Reverberation can be modeled as virtual "image sources" in virtual rooms, denoted by the smaller asterisks outside the boundaries of the physical room. Each image source is the result of reflecting the physical source about one or more physical or virtual walls.



(a) Reverberant room impulse response



(b) Log magnitude reverberant response

Figure 2-4: An example reverberant response (a) and its log magnitude (b). The tail of the response decays exponentially, as can be seen by the linear decrease in average log magnitude.

PHAT weighting is the ML weighting for the case of constant SNR across frequency.

Now let us look back at reverberation and see why it might make sense to assume constant SNR across frequency. We model reverberation as the result of applying a filter, such as the one shown in Figure 2-4(a) to the source signal

$$x_i(t) = h_i(t) * s(t) + n_i(t) \tag{2.36}$$

where $h_i(t)$ is the linear time-invariant (LTI) system representing the effects of the acoustic environment. We can decompose this into a direct path component and a reverberant component and rewrite it as

$$\begin{aligned} x_i(t) &= h_{i_{direct}}(t) * s(t) + h_{i_{reverb}}(t) * s(t) + n_i(t) && (2.37) \\ h_{i_{direct}}(t) &= \frac{1}{r_i}\delta(t - \frac{r_i}{v}) && (2.38) \\ h_{i_{reverb}}(t) &= h_i(t) - h_{i_{direct}}(t) && (2.39) \end{aligned}$$

where $h_{i_{direct}}(t)$ is the direct path component, corresponding to the earliest peak in Figure 2-4(a), and $h_{i_{reverb}}(t)$ is everything else, consisting of reflections coming from many different directions. Since this reverberant component appears to be coming from directions other than the source direction, it is effectively noise, and we can define a frequency-specific equivalent SNR (similar to [18]):

$$SNR_{eq}(f) = \frac{H_{i_{direct}}(f)S(f)}{H_{i_{reverb}}(f)S(f) + N_i(f)} \tag{2.40}$$

When $N_i(f) \ll H_{i_{reverb}}(f)S(f)$, this is approximately $H_{i_{direct}}(f)/H_{i_{reverb}}(f)$, which does not depend on the magnitude of the signal, $S(f)$. So if the reverberation is equally strong at all frequencies, then the effective SNR is the same at all frequencies, and this provides an intuitive justification for GCC-PHAT's constant weighting.

Of course, this all depends on whether we are justified in treating the reverberant component as "noise" that meets the assumptions of the GCC derivation, namely that

the noise is uncorrelated with the target signal and across microphones. This seems problematic since our "noise" is just a filtered version of our signal and is therefore correlated with it.

Gustafsson *et al.* [38] show us a way out of this predicament. Instead of defining uncorrelatedness as a statistical expectation across time for a fixed source-microphone configuration, we can define it as an expectation across room configurations in which we randomize the source and microphone positions. They show, using results on the statistics of room acoustics from [83], that these uncorrelatednesses are approximately true for high enough frequencies (in relation to the size of the room) as long as the microphones are far enough apart and as long as the microphones are not too close to any large objects.

So, in summary, the GCC-PHAT weighting is approximately the GCC ML weighting for stationary signals in environments with negligible additive noise and reverberation that is equally strong across frequency. (However, reverberation strength varies with frequency in many environments because the sound absorption of many common building materials varies with frequency [58].)

2.3 Psychoacoustic background

In this section, we briefly review human sound localization, with emphasis on the precedence effect. Because of the interaction of the rich natural environment, the complexity of the human brain, and the physiological limitations of the auditory system, human sound localization is much harder to analyze than the idealized systems we looked at in the previous sections. For more information on human sound localization, Blauert [10] seems to be the most complete single source. Yost and Gourevitch [94] have edited together a number of useful contributions, and [59] is a good recent review of the precedence effect.

2.3.1 Localization cues

Humans are subject to the same fundamental physical and signal processing limitations as any other system, so it is not surprising that, as predicted by in the analysis of Section 2.1, time-delay-of-arrival between the two ears is one of the primary cues that people use to localize sounds.

Humans extract two types of time-delay information from the signal. The first is time delay derived from the fine structure of the signal. This comes from phase differences in narrow signal bands and is much like the phase differences that were our focus in the mathematical analysis. Humans use this cue only up to roughly 1.5 kHz. At higher frequencies, narrowband phase difference cues become increasingly ambiguous. The other type of time delay information in the human auditory system is time delay between signal envelopes. (Roughly speaking, signal envelopes are rectified and low-pass filtered versions of the signals. They capture overall energy variation and ignore the detailed structure of the signal.) Because they are low-pass filtered, even the envelopes of sounds that contain only frequencies above 1.5 kHz are not subject to the ambiguities present in the interaural phase differences. In addition, the human auditory system does not transmit phase information to the brain for frequencies above 4-5 kHz [35], so at these very high frequencies, envelope time delays are the only possible source of binaural time delay information.

Humans also use differences in sound amplitude at the two ears to determine source location. People have heads between their two ears for a variety of reasons, but the relevant reason for the present discussion is that the head serves to "shadow" each ear from sounds originating on the other side of the head. This "shadowing" is most significant for wavelengths small in comparison to the size of the head. Conveniently, this is the opposite of the case with phase differences, where short wavelengths are less useful because there are multiple feasible delays corresponding to the same phase difference. In normal listening situations, level (amplitude) difference cues are useful for sounds with energy above 1.5 kHz.

Level differences and both types of time delay cues are binaural cues, in that they

use the differences between the signals at the two ears to localize the sound source. These binaural cues are subject to certain ambiguities, however. For example, any sounds originating from the mid-sagittal plane (the plane perpendicular to the line joining the ears and half way between them) will travel an exactly symmetric path to each ear and will result in exactly the same signal being received. (This assumes anechoic conditions. In reverberant conditions, the signals at the two ears may be different, but the differences will not be useful for localization.) Since the signals are the same, there are no binaural differences to use as localization cues.

For cases such as this where binaural cues are ambiguous, the human auditory system uses monaural spectral cues. Depending on the source location of a sound, interactions of the travelling wave with the body, head, and outer ears will filter the signal in specific ways. Thus if the original spectrum emitted by the source is known, its direction can be estimated based on differences between this original spectrum and the spectrum of the signal received at the ears. Because of the asymmetries of the outer ears, head, and body, these cues can disambiguate different source locations in the mid-sagittal plane, so they work to complement the binaural cues.

Humans appear to make good use of the localization cues available to them, for example by using interaural time delays and interaural level differences in complementary frequency ranges to ensure that sounds at all frequencies can be localized. This suggests that we should look to psychoacoustics for ways to improve the performance of our automated source localization techniques, particularly as they apply to the complex signals and reverberant environments in which people normally operate. The precedence effect suggests a way to do exactly that.

2.3.2 The precedence effect

The precedence effect, also known as the "law of the first wavefront," is the psychoacoustic effect in which the apparent location of a sound is influenced most strongly by the localization cues from the initial onset of the sound [59, 96]. For example, when human listeners report the location of a rapid sequence of clicks, they tend to report the location of the initial click even if later clicks in the sequence came from other

directions [59]. It has been argued that the precedence effect improves people's ability to localize sounds in reverberant environments [95] because direct path sound arrives before any correlated reflections, so initial onsets will tend to be less corrupted by reverberation than subsequent sounds. The generality of this argument suggests that other animals should also exhibit the precedence effect, and evidence for the effect has been found in cats, dogs, rats, owls, and crickets [59].

Although the basic utility of the precedence effect seems straightforward, the details are not clear. The notion of an "onset" is imprecise, although progress has been made in [87] in determining the time scales over which the precedence effect operates for click trains, and [76] shows the effect of onset duration on the ability to localize narrowband sounds. In addition, most studies have focused on stimuli such as click trains or noise bursts, and it is not obvious how to apply their findings to more natural sounds. For example, the effect is strongest in click pairs for inter-click intervals of roughly 2-10ms [59]. Shorter inter-click delays result in "summing localization," where a single click at some intermediate location is perceived. Longer inter-click intervals result in the the perception of two clicks at two separate locations.

Studies on human infants (reviewed in [59]) found no evidence of the precedence effect, and studies on young children have found the effect to be much smaller. Studies on puppies [4] have shown that the precedence effect develops significantly after the basic ability to localize sounds. This suggests that the precedence effect may be learned during childhood, although maturation of neural pathways, even in the absence of direct experience in reverberant environments, could also cause this gradual development of the effect.

As with most psychoacoustic phenomena, there are some subtleties. For example, in the "Clifton effect" [20], the precedence effect can be temporarily suppressed by suddenly swapping the locations of the leading and lagging clicks in a click-pair experiment. Another subtlety is that if several click pairs are played sequentially, the dominance of the initial click in each pair will increase for the later pairs. The reasons for these behaviors are not well understood, but they may be a result of the brain doing some sort of online learning of its acoustic environment.

A number of computational models of the precedence effect have been proposed. In [96], Zurek proposed a high-level conceptual model of the precedence effect without precisely specifying the details of the model. He modeled the precedence effect as a time-dependent weighting of raw localization cues. Specifically, his weighting took the raw audio as input and consisted of an "onset detector" with output generated by an inhibition function. Zurek's high-level model was subsequently implemented and evaluated by Martin [62].

Lindemann [56, 57] presents a cross-correlation-based model of auditory localization, subsequently extended by Gaik [30], that includes an inhibition component that can model many aspects of the precedence effect. Lindemann's model has many parameters whose values were chosen to accurately model human localization performance. Huang *et al.* [45] and Bechler [6] present more engineering-oriented models of the precedence effect and apply them to source localization. However, their approaches make all-or-none decisions about each localization cue. Also, Huang *et al.* base their time delay estimates on differences between zero-crossing times instead of finding the maximum of a cross-correlation function. Recently, Faller and Merimaa [29] presented a model that uses estimated interaural coherence values to predict which time instants in a reverberated signal contain the best localization cues. They model many of the aspects of the precedence effect using these interaural coherence values, but their model does not explain why some steady-state sounds with high coherence are suppressed or why sounds originating in the median sagittal plane, which are perfectly coherent, can still elicit the precedence effect as shown in [60].

The model that we will describe in the next chapter can be viewed as a specific implementation of a model similar to Zurek's. However, our goal is not to faithfully model the human auditory system but rather to find a weighting function for the GCC framework that will accurately localize speech in reverberant environments. Because of this difference in approach, we do not incorporate elements such as psychoacoustically inspired filter banks or neural transduction models, and we do not try to model details such as the Clifton effect. Instead we focus on predicting the reliability of localization cues derived from a simple spectrogram representation. In

contrast to other precedenced-based approaches, our approach relates directly to the GCC framework, which is the optimal TDOA estimator (under a set of assumptions enumerated in [52]) and provides a principled way to integrate localization cues across time and frequency. In contrast to Faller and Merimaa, who make use of interaural coherence, we predict localization precision based on monaural cues, which we know are psychoacoustically relevant from [60]. In contrast to [6, 29, 45], our technique will not make all-or-nothing decisions about whether to use a localization cue. Instead, we will use a continuous measure of cue reliability, which makes more sense from a signal processing perspective and is also more consistent with the psychoacoustic data [96, p. 95].

Chapter 3

Localization Algorithm

This chapter describes our localization algorithm, which takes the form of a new GCC weighting function.

Based on the theoretical justifications in Chapter 2, we know that even in reverberant environments, if we knew the phase estimate error variance across frequency, we could use generalized cross-correlation to do approximately optimal time delay estimation. Our goal, then, will be to estimate this error variance.

Since phase error variance is related to SNR, one way to estimate error variance is to estimate the noise power during silence and the signal power during speech activity, and to use these estimates to calculate the phase error variance [12]. Because the estimate of the "signal" power will also typically include reverberant energy (which is effectively noise for the purpose of delay estimation), however, this estimate performs inadequately [13, 22]. Another problem with this approach is that getting accurate signal and noise power estimates requires the use of a long observation window. However, speech is nonstationary, so long observation windows will average out potentially important changes in signal power.

We take a different approach. In the end, we do not care about the particular signal power or noise power. We only care about accurately predicting the phase error variance, and any observable variables that have some dependence on the phase error variance can be used to (imperfectly) predict it. The precedence effect suggests that people use onsets in the input audio signal to decide which localization cues to em-

phasize, so we will design a system that can capture this relationship and potentially other relationships that may exist. Our new goal, then, is to learn features observable in the received audio that predict the reliability of associated intermicrophone phases. (We choose not to explicitly model the precedence effect because our more general model is easy to optimize and can potentially model other relationships.)

This goal is intuitively appealing but still imprecise. We need to say what we mean by "learning," and we need to pick a set of features and a functional form for our predictor. "Learning" in this thesis will be solving a regression problem using labelled training data. In our implementation we choose to use linear regression to find a mapping from the audio log-spectrogram to the localization log-precision, which we define to be the logarithm of the reciprocal of the empirical TDOA mean-squared error. We have no fundamental theoretical justification for these choices, but we have several practical justifications:

1. Log-spectra and Log-precisions range from $-\infty$ to ∞. Without the logarithm, they would range from 0 to ∞. Unconstrained linear functions are capable of generating negative values, which are incompatible with a non-log domain. (It is possible to constrain the output to be positive using nonnegative matrix factorization [55], but this is more computationally intensive and implies a parts-based representation [54] that does not necessarily apply to our problem.)

2. In a fixed, purely reverberant environment with no additive noise, changing the overall loudness of the source signal should not change the phase error since in this case all of the "noise" is reverberant, and as the source signal gets louder, the reverberant noise will also get proportionally louder (as described in Section 2.2.2). Scaling a signal's amplitude corresponds to an additive shift (a change in the DC component) in the log domain. Linear mappings whose coefficients sum to zero will not pass this DC shift and will therefore be invariant to overall signal scaling, so they are capable of capturing this invariance.

3. Linear functions are computationally efficient to train and to apply. In the absence of an argument against them, they should be one of the first things

tried. (We have done some small-scale experiments using quadratic terms or using quadratic or Gaussian kernels, but so far have found only negligible improvement with these techniques.)

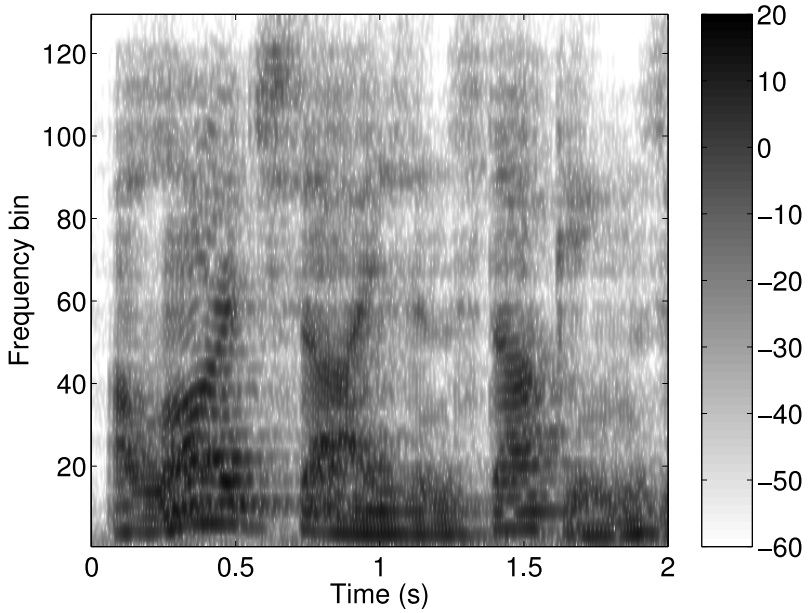
4. Log-spectrum-based representations (particularly the mel-frequency cepstral representation) are among the most successful representations used in automated speech recognition [46]. They have also been used successfully for speech denoising [26]. Linear functions applied to log-spectral representations have also proven useful. For example, delta cepstral features often improve automated speech recognition performance, and linear operations have also been used to successfully denoise cepstral coefficients [27].

Now that we have chosen our input representation, our output representation, and our regression technique, we just need some labelled training data. An example of such training data, consisting of a reverberated speech log-spectrogram as input and a corresponding time-frequency map of log-precision as output, is shown in Figure 3-1. The next section describes how this pair was generated.

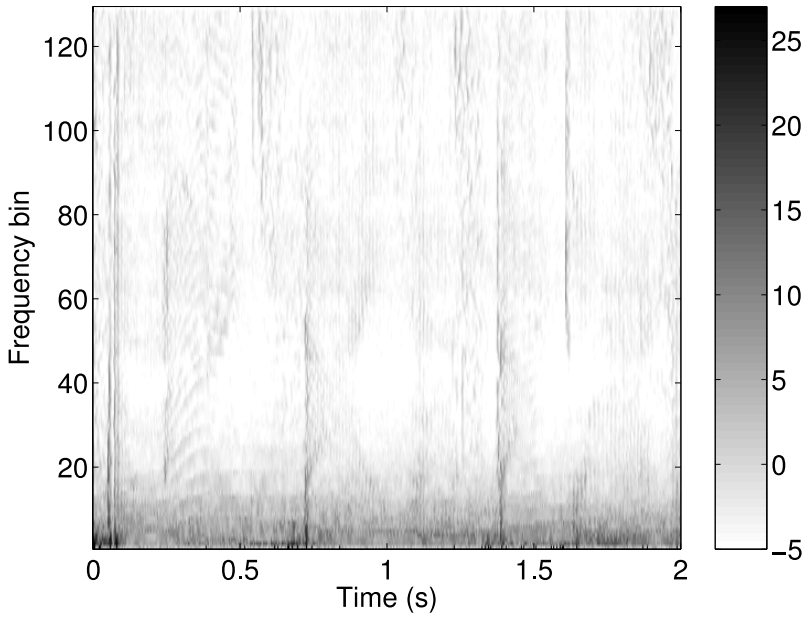
3.1 Corpus generation

Our training corpus consists of reverberant speech spectrograms and time-aligned time-frequency maps of phase log-precision for a large set of spoken utterances, as shown in Figure 3-1. The spectrograms are computed using standard methods, for example as in [46, p. 281]. The time-frequency localization precisions, which we will refer to as precision-grams, are specific to our technique, however. To generate a precision-gram, we collect N_r realizations of each utterance, each with the speech source and microphones in different locations. We then calculate the empirical localization precision over all realizations.

More formally, we start with a single speech signal, $x(t)$, and randomly generate N_r simulated room configurations. For our experiments, microphone and source location and room size and shape vary across configurations. We represent these room



(a) Speech spectrogram



(b) Localization precision map

Figure 3-1: Example data from our training corpus. Figure 3-1(a) is a spectrogram of the reverberant speech (a male voice saying "A large size in stockings...") received at one of the microphones in the array. Figure 3-1(b) is the corresponding map of the empirical localization precision (in dB) for each time-frequency bin. Note that sudden onsets in the spectrogram (a), such as those at 0.07, 0.7, and 1.4 seconds, correspond to time-frequency regions with high localization precision in (b).

configurations as filters $H_j(i,t)$, where $j \in \{1 \ldots N_r\}$ represents the room realization and $i \in \{1,2\}$ represents the i^{th} microphone signal. Passing $x(t)$ through $H_j(i,t)$ and adding a noise signal $n_j(i,t)$ yields $y_j(i,t)$, a set of reverberated speech signals. We then compute spectrograms of $y_j(i,t)$ with window size N_w, overlap N_o, and FFT length N_f, yielding complex spectrograms $s_j(i,u,f)$, where frame index u replaces the time index t, and frequency index f is added. We then calculate the cross-power spectrum phase,

$$\theta_j(u,f) = \angle \frac{s_j(1,u,f)}{s_j(2,u,f)} \tag{3.1}$$

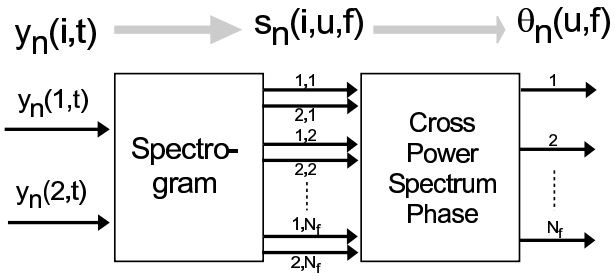
for each frame and frequency bin. (The cross-power spectrum phase will range from $-\pi$ to π.) Finally, we calculate $\tilde{\sigma}^2(u,f) = \frac{1}{N_r}\sum_{j=1}^{N_r}(\theta_j(u,f) - \theta_{j_{true}}(u,f))^2$, the localization (wrapped phase) error variance, and $\tilde{q}(u,f) = -10 * log_{10}(\tilde{\sigma}^2(u,f))$, the localization precision (in dB). We determine $\theta_{j_{true}}(u,f)$ from the assumed known source to array geometry in our training set. Figure 3-2(a) shows a block diagram describing these calculations.

By calculating only these variances without any cross-covariances we implicitly assume that localization errors in different time-frequency regions are uncorrelated. Gustafsson et al. [38] showed that this is approximately true in reverberant environments, and we have found this assumption to work well in practice.

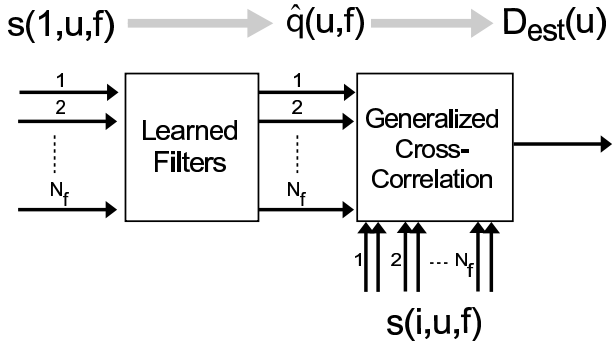
3.2 Filter learning

Once we have collected a training corpus, we use ridge regression [32] to learn FIR filters that estimate the localization precision (in dB) from the reverberated spectrogram (in dB). In this thesis, we examine two different forms for these filters.

In the first case, which we call a narrowband mapping, we learn a separate FIR filter from each frequency band in the spectrogram to the corresponding frequency band in the localization precision output as shown schematically in Figure 3-3(a). In the second case, which we call a broadband mapping, we learn a separate FIR filter for each band of the localization precision output, but in each case the input comes

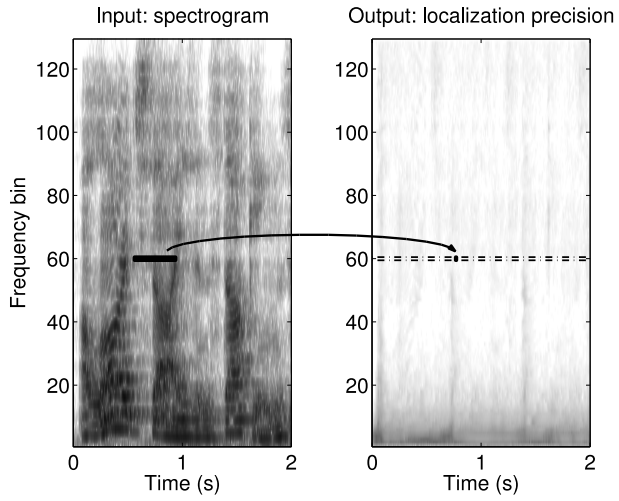


(a) Phase calculation during training

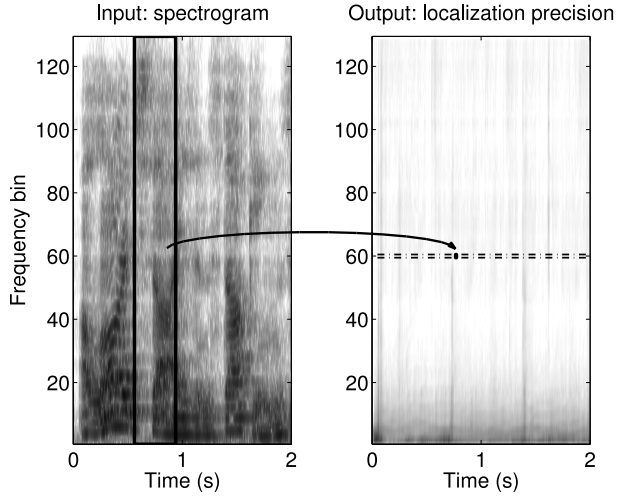


(b) TDOA calculation during testing

Figure 3-2: (a) shows the procedure for calculating the cross-power spectrum phase used during training. (b) shows the procedure for using our estimated precision map to calculate TDOA during testing.



(a) Narrowband precision calculation



(b) Broadband precision calculation

Figure 3-3: An illustration of the narrowband and broadband mappings for frequency band 60. In (a) an FIR filter estimates the localization precision as a function of spectrogram bin 60. In (b) an FIR filter estimates the localization precision as a function of all spectrogram bins.

from all frequencies of the input spectrogram. This case is shown schematically in Figure 3-3(b). Our motivation for examining the narrowband case is that, for the case of stationary signals (and under the assumption of spectrogram windows that are much larger than the coherence time of the signal), each frequency band is uncorrelated with all other frequency bands, and thus the narrowband mapping should be sufficient in this case. Although speech is nonstationary, this narrowband mapping provides a useful baseline against which to compare. Additionally, in [76], the precedence effect was demonstrated with narrowband sounds, where the onset rate of a sinusoidal tone affected the ability to localize that tone, which is exactly the relationship that our narrowband mapping can express. The broadband mapping subsumes the narrowband mapping and should be able to capture cross-frequency dependencies that may arise from the nonstationarity of speech. Such cross-frequency dependencies have been observed in psychoacoustic studies of the precedence effect [84].

For the narrowband mapping with causal length l_c and anticausal length l_{ac}, we solve N_f regularized linear least-squares problems of the form $\mathbf{z_f} = \mathbf{A_f b_f}$, $f \in \{1 \dots N_f\}$ where

$$\mathbf{z_f} = \begin{pmatrix} \dots & \tilde{q}(u,f) & \tilde{q}(u+1,f) & \dots \end{pmatrix}^{\mathsf{T}}$$

$$\mathbf{A_f} = \begin{pmatrix} \vdots & \vdots & \vdots & \vdots & \vdots \\ s(1,u-l_c,f) & s(1,u+1-l_c,f) & \dots & s(1,u+l_{ac},f) & 1 \\ s(1,u+1-l_c,f) & s(1,u+2-l_c,f) & \dots & s(1,u+1+l_{ac},f) & 1 \\ s(1,u+2-l_c,f) & s(1,u+3-l_c,f) & \dots & s(1,u+2+l_{ac},f) & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \tag{3.2}$$

and $\mathbf{b_f}$ is an FIR filter with $(l_c + l_{ac} + 1)$ taps stacked with a DC component. $s(1, u - l_c, f)$ is the spectrogram from the first channel of a randomly chosen room configuration.

For the broadband mapping, we solve N_f regularized linear least-squares problems of the form $\mathbf{z_f} = \mathbf{A_f b_f}$, where

$$\mathbf{z_f} = \left(\ldots \; \tilde{q}(u,f) \; \tilde{q}(u+1,f) \; \ldots \right)^{\mathsf{T}}$$

$$\mathbf{A_f} = \begin{pmatrix} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s(1,u-l_c,1) & \ldots & s(1,u+l_{ac},1) & s(1,u-l_c,2) & \ldots & s(1,u+l_{ac},2) & \ldots & s(1,u+l_{ac},N_f) & 1 \\ s(1,u+1-l_c,1) & \ldots & s(1,u+1+l_{ac},1) & s(1,u+1-l_c,2) & \ldots & s(1,u+1+l_{ac},2) & \ldots & s(1,u+1+l_{ac},N_f) & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \tag{3.3}$$

and $\mathbf{b_f}$ is an FIR filter with $(l_c+l_{ac}+1)*N_f$ taps stacked with a DC component. For both types of mapping, we solve these systems using ridge regression by minimizing

$$||\mathbf{z_f} - \mathbf{A_f}\mathbf{b_f}||^{\mathbf{2}} + \lambda||\mathbf{b_f}||^{\mathbf{2}} \tag{3.4}$$

with respect to $\mathbf{b_f}$. The regularizing parameter λ is set using a validation data set, and results were found to be relatively insensitive to the particular choice of λ.

3.3 Applying the filters

We apply filters $\mathbf{b_f}$ to spectrogram $s_n(1,u,f)$ yielding $\hat{q}(u,f)$. (In most cases in this thesis we will use the tilde accent mark to indicate a sample-based estimate, the caret accent mark to indicate the result of applying one of our mappings, and no accent mark to indicate the "true" underlying value, as in $\tilde{\sigma}^{-2}(u,f)$, $\hat{\sigma}^{-2}(u,f)$, and $\sigma^{-2}(u,f)$, respectively.) We then use this estimated log-precision to create a GCC weighting for each frame, which we define to be

$$\Psi(u,f) \triangleq \frac{\hat{\sigma}^{-2}(u,f)}{|G_{x_1x_2}(u,f)|} \tag{3.5}$$

$$\hat{\sigma}^{-2}(u,f) = 10^{\frac{\hat{q}(u,f)}{10}} \tag{3.6}$$

Here, we are using $\hat{\sigma}^{-2}(u,f)$ as an approximation to the $B(f)$ term in Equation 2.12. This is justified because we showed in Equation 2.27 that $B(f)$ is approximately the phase error variance. Note also that the phase transform is equivalent to setting

$\hat{\sigma}^{-2}(u, f) = 1$. (In our discussion of the theory in Chapter 2, we focused on a delay estimate based on a single windowed observation, so $\Psi(f)$ depended only on frequency. We are now computing a generalized cross-correlation function for each spectrogram frame, so we have a corresponding weighting function for each frame, and $\Psi(u, f)$ now depends on both frame index and frequency.)

When applying this technique to localization, the only computational costs (beyond the basic TDOA calculations) are of applying a set of short FIR filters to that spectrogram. Because our training set encompasses many different acoustic environments, the mappings that we learn do not depend strongly on the detailed structure of the reverberation, and our technique is robust to changes in the acoustic environment.

3.4 Related work

In addition to the computational models of the precedence effect mentioned in Chapter 2, there has been other work on creating statistical models or finding confidence metrics for localization cues.

Bechler and Kroschel [5] propose the use of two confidence metrics for evaluating cross-correlation-based time-delay estimates. In both cases, their intuition is that cross-correlation waveforms with more well-defined peaks should lead to better time delay estimates. Their first criterion is "maximum peak height," and their second is "peak ratio," or the ratio between the tallest and second-tallest peaks in the cross-correlation waveform. They show that the number of outlier time delay estimates decreases monotonically with each of these criteria. However, they do not show that by applying these reliability estimates they can subsequently improve localization over standard techniques such as GCC-PHAT. In addition, each of their reliability criteria applies to a single frame of data across all frequencies, whereas our technique gives both time- and frequency-dependent reliability.

Roman [79, 80] and Nix [66, 68] have both recently done interesting work modeling the statistics of localization cues from empirical observations, but with a more psychoacoustically-faithful focus. Both learn statistical models for both interaural

level differences and interaural time differences for microphones located in the ears of a real (for Nix) or KEMAR mannequin (for Roman) head. When a head and ears are present, the relationship between the source location and these localization cues becomes much more complicated, necessitating models based on empirical data to achieve even reasonable baseline performance. Both Roman and Nix effectively marginalize out any time-dependence in the distributions of the localization cues, so they are incapable of capturing the precedence effect.

In contrast to our work, they provide no direct link between their confidence metrics and the optimal signal processing approaches to time delay estimation, and they make no quantitative comparison to standard techniques such as GCC-PHAT. We have also avoided the use of a "head," and therefore the use of interaural level differences, because the presence of a head may not be desirable in many practical application scenarios and because in practice as long as timed delay cues are available, they tend to be more reliable than level-difference cues.

Chapter 4

Localization Experiments

This chapter addresses the following questions:

1. How well does time delay estimation based on our learned localization precision mappings work in comparison to existing time delay estimation techniques?

2. Are the learned localization precision mappings consistent with the precedence effect?

To simplify data collection and allow for comparison under repeatable and precisely characterizable acoustics, the bulk of the experiments were done on synthetic data. Additional experiments were done on real data to demonstrate that our technique does not depend on any peculiarities or artifacts in the synthetic data.

4.1 Experimental scenario

These experiments test the ability of a two-element microphone array to localize a single speaker in a reverberant environment. No competing speaker is present, but stationary background noise (or slowly varying, approximately stationary for the data from real rooms) is present.

Room ID	Dimensions (meters)	RT_{60} (seconds)
A	$4 \times 7 \times 2.8$	0.1
B	$4 \times 7 \times 2.8$	0.2
C	$4 \times 7 \times 2.8$	0.4
D	$8 \times 14 \times 5.6$	0.8
E	$16 \times 28 \times 11.2$	1.6

Table 4.1: Room dimensions and reverberation times for synthetic rooms.

Room ID	Dimensions (meters)	RT_{60} (seconds)
F	$4.9 \times 7.3 \times 2.7$	0.6
G	$3.2 \times 6.2 \times 3.0$	0.4
H	$12.8 \times 6.7 \times 10.7$	0.8

Table 4.2: Room dimensions and reverberation times for real rooms. Rooms F, G, and H are MIT campus rooms 32-D507, 32-D514, and 32-D463, respectively.

4.1.1 Synthetic data set

The primary goal of the experiments on synthetic data is to explore time delay performance as a function of background noise level and reverberation time. To do this, we simulated 5 rooms with reverberation times ranging from 100 ms to 1600 ms. Room dimensions and reverberation times are listed in Table 4.1. Rooms A through C all have the dimensions of a typical office or small conference room and differ only in the absorption coefficients of their simulated walls. Rooms D and E have walls that are as absorptive as room C's and achieve their longer reverberation times because of larger dimensions, which are representative of larger conference rooms or lecture halls.

Within each room, 60 distinct source-microphone configurations were simulated for training, and a separate set of 60 distinct source-microphone configurations were simulated for testing. The image method [3] with sub-sample interpolation to ensure accurate inter-microphone phase [73] was used to simulate reverberation for both testing and training. The simulated inter-microphone distance was 20 centimeters.

The speech material was taken from the TIMIT corpus [31], a multi-speaker corpus recorded with a close-talking microphone. For testing, we use the TIMIT core test set, which consists of sixteen males and eight females, each providing roughly 30 seconds of speech. For training, we use a randomly selected eight males and eight females

from the TIMIT training set, which does not overlap with the TIMIT test set. The TIMIT data was originally recorded at 16 kHz sampling rate, and we downsample to 8 kHz before training and before applying any of the localization techniques. To test robustness to noise, varying levels of stationary Gaussian white noise were added to the reverberant speech signals. The noise signals are uncorrelated across microphones.

We train a single set of filters across all noise levels, reverberation times, and speakers, and apply this single set of filters to all testing conditions.

4.1.2 Real data set

Real data was collected in three rooms in MIT's Stata Center (details in Table 4.2). Room F is a medium-size meeting room. Room G is a typical office, and room H is a large conference room. In each room, each of two speakers stood in one of twelve distinct locations. The speaker locations did not repeat, so between the two speakers in each room, 24 total source-microphone configurations were tested in each room. In each configuration, approximately one minute of speech was recorded. Data was collected at 44.1 kHz and subsequently downsampled to 8 kHz. The intermicrophone spacing was 37.5 cm. Figure 4-1 shows the data acquisition setup used for these experiments. The upper pair of microphones was used for all experiments.

Background noise in each case was the ambient noise that was present at the time. In all cases this consisted of at least the fan noise from the laptop to which the microphone array was connected and some slowly varying room ventilation noise.

For the real data experiments, results for each room were obtained by training on data from the other two rooms.

4.2 Description of Compared Techniques

We compare the performance of eight time-delay estimators on our synthetic data. All but one technique fit in the GCC framework, and each technique's corresponding weighting function is listed in Table 4.3. The first two are oracle-based methods that provide upper bounds on performance but cannot be applied in practice because

Figure 4-1: The microphone array setup used for all experiments on real data. Microphones (highlighted by grey circles) are held in a plastic frame surrounding the laptop screen. The top two microphones were used in all experiments.

Technique	GCC weighting function $\Psi(f)$
GCC-ML	$\frac{\lvert\tilde{\gamma}_{x_1x_2}(f)\rvert^2}{\lvert G_{x_1x_2}(f)\rvert[1-\lvert\tilde{\gamma}_{x_1x_2}(f)\rvert^2]}$
Emp. Prec.	$\frac{\tilde{\sigma}^{-2}(f)}{\lvert G_{x_1x_2}(f)\rvert}$
Broadband	$\frac{\hat{\sigma}^{-2}_{broadband}(f)}{\lvert G_{x_1x_2}(f)\rvert}$
Narrowband	$\frac{\hat{\sigma}^{-2}_{narrowband}(f)}{\lvert G_{x_1x_2}(f)\rvert}$
Proportional	$\frac{\hat{\sigma}^{-2}_{proportional}(f)}{\lvert G_{x_1x_2}(f)\rvert}$
GCC-PHAT	$\frac{1}{\lvert G_{x_1x_2}(f)\rvert}$
Cross Correlation	1
Benesty	N/A

Table 4.3: Generalized cross-correlation weighting functions for each technique.

they depend on information that is available in simulation only. The next three are variants of our localization precision estimation technique. The following two are standard generalized cross-correlation (GCC) variants, and the final one is an adaptive eigenanalysis technique developed by Benesty [7, 47]. We compare the six non-oracle techniques in the experiments in real rooms.

We do not compare to any of the computational models of the precedence effect mentioned in Chapter 2 because we are not aware of any of those models demonstrating performance superior to standard techniques such as GCC-PHAT under typical operating conditions. Those models serve primarily as tools to better understand human psychoacoustics, not as practical localization techniques.

"GCC-ML" is the weighting for the maximum-likelihood time-delay estimator described in [52]. This method requires knowledge of $|\gamma_{x_1x_2}(f)|^2$, the inter-microphone magnitude-squared coherence function. Based on Equation 2.11, $|\gamma_{x_1x_2}(f)|^2$ is

$$|\gamma_{x_1x_2}(f)|^2 = \frac{|G_{x_1x_2}(f)|^2}{G_{x_1x_1}(f)G_{x_2x_2}(f)} \tag{4.1}$$

where $G_{x_i x_j}(f)$ is the cross-spectrum of microphone channels i and j (the auto-spectrum for the case $i = j$). The speech signals to which we apply this technique are nonstationary, so we must use a short-term (time-windowed) version of the cross-spectrum. For this evaluation, we assume that the signal and noise are known and sum to the observed microphone signal

$$x_i(t) = s_{i_{direct}}(t) + n_i(t) \tag{4.2}$$

where $s_{i_{direct}}(t)$ is the direct-path signal from the speaker to microphone i and $n_i(t)$ is the noise, including any uncorrelated additive noise in addition to reverberation. Using our perfect knowledge of the direct path signal component within our simulation, we can calculate a short-term magnitude-squared coherence as

$$|\tilde{\gamma}_{x_1 x_2}(f)|^2 = \frac{[S_1(f)S_2^*(f)]^2}{[S_1(f)S_1^*(f) + N_1(f)N_1^*(f)][S_2(f)S_2^*(f) + N_2(f)N_2^*(f)]} \tag{4.3}$$

where $X_i(f)$ and $N_i(f)$ are Fourier-transforms of time-windowed segments of $x_i(t)$ and $n_i(t)$. This definition of $|\tilde{\gamma}_{x_1 x_2}(f)|^2$ amounts to assuming that signal and noise are uncorrelated and estimating cross-spectral densities in Equation 4.1 based on only the current time window.

For stationary signals, it is possible to calculate a non-oracle estimate of the intermicrophone coherence by segmenting the signal into a number of finite-length observations and computing phase variation across these segments [17], and this coherence estimate can then be used to estimate the GCC-ML weighting. Brandstein et al. [12] evaluated the use of such a technique for speech source localization. Because speech is nonstationary, coherence estimates must be based on very short segments, and they found that coherence estimates from such a small amount of data led to poor performance. For these reasons, we did not include such a non-oracle estimated GCC-ML weighting in our evaluation.

"Empirical precision," the second oracle-based technique, should be an upper bound on the performance of all of our localization precision mapping techniques.

It directly uses the empirically determined (ground truth) precision of each time-frequency region in the test set. To the extent that our weightings underperform the true precision it is presumably due to their inability to perfectly reconstruct this ground truth precision.

The next three techniques are variations on the localization precision estimation technique. "Broadband" and "Narrowband" are the mappings described in Section 3.2. "Proportional" is a simple special case of the narrowband filter using only one tap. This "proportional" mapping could express the simple relationship in which localization cues are weighted proportionally to the local signal power, but it cannot capture more complicated relationships.

"GCC-PHAT" is the phase transform, and it corresponds to uniformly weighting the localization cues in each time-frequency region (setting $\hat{\sigma}^{-2}(f) = 1$). "Cross correlation" is a simple cross correlation with no weighting applied.

"Benesty" is the adaptive eigenanalysis technique described in [7, 47]. Here we briefly outline the technique and contrast it with GCC-based techniques.

4.2.1 Adaptive eigenanalysis time delay estimation technique

Although they are often applied in reverberant environments, GCC-based techniques are typically motivated with the assumption that noise in one channel is uncorrelated with noise in the other channel and with the target signal. Under these assumptions, the goal of the weighting function is to emphasize frequencies with high SNR, which will bring the peak of the cross-correlation waveform out of the noise.

Benesty's technique is explicitly formulated to take reverberation into account. He starts from the model

$$x_i(t) = h_i(t) * s(t) + b_i(t) \tag{4.4}$$

where $s(t)$ is the target signal at its source, $h_i(t)$ is the transfer function from the source to microphone i, and $b_i(t)$ is an uncorrelated noise signal. To simplify the presentation of Benesty's technique, we will assume that $b_i(t) = 0$. In the noiseless

case, based on Equation 4.4, the following relationship holds:

$$h_1(t) * x_2(t) = h_1(t) * h_2(t) * s(t) = h_2(t) * x_1(t) \tag{4.5}$$

The $h_i(t)$ are unknown, but Benesty uses the relationship in Equation 4.5 to find them using a discrete-time adaptive filtering problem by first defining

$$\begin{aligned}
\mathbf{x_i}(n) &= [x_i(n), x_i(n-1), \cdots, x_i(n-M+1)]^T & (4.6)\\
\mathbf{h_i} &= [h_i(0), h_i(1), \cdots, h_i(M-1)]^T & (4.7)\\
\mathbf{x}(n) &= [x_1^T(n), x_2^T(n)]^T & (4.8)\\
\mathbf{u} &= [h_2^T, -h_1^T]^T & (4.9)
\end{aligned}$$

where M is the length of the adaptive filter.

He then uses standard adaptive filtering techniques [61] to find $\mathbf{u}$ based on

$$\mathbf{x^T}(n)\mathbf{u} = \mathbf{x_1^T}(n)\mathbf{h_2} - \mathbf{x_2^T}(n)\mathbf{h_1} = 0 \tag{4.10}$$

When found subject to appropriate constraints, $\mathbf{u}$ consists of estimates of the two source-microphone transfer functions. Benesty then finds the locations of the largest peaks in the two estimated transfer functions and defines his delay estimate to be the difference in lags between the peak locations. Benesty demonstrated that his technique had lower time delay errors than GCC-PHAT on 5 second-long audio segments in a wide range of reverberation times.

4.3 Performance results

We evaluate all techniques on 500 ms segments of audio during which the source remains motionless. 500 ms is enough time for a speaker to utter two to three syllables, so the ability to localize these segments implies the ability to localize all but the shortest conversational utterances. In addition, studies of human localization perfor-

mance have shown that performance improves as segment length increases, but that performance on bandlimited noise begins to asymptote around 500 ms duration [10, p. 156]. For our results, these 500 ms segments come from contiguous nonoverlapping segments of continuous speech. Because of pauses between words or phrases, some of these segments will consist wholly or partially of silence.

For all of the GCC techniques, we use a 150-sample (18.75 ms) window and 30-sample (3.75 ms) step size (at the 8kHz sampling rate) when computing cross-correlation waveforms. For the "Benesty" technique, we choose the adaptive filter length, M, to be 150 samples to match our GCC window size. Benesty's technique also requires an adaptive update rate parameter μ. We use $\mu = 0.003$, which is the value used in [47] and which we also found to give good performance.

When estimating a time delay by finding a cross-correlation peak location, there are two types of errors, local errors and gross errors [48, 49, 89]. Local errors occur when the noise perturbs the location of the peak. Gross errors occur when noise causes the wrong peak to be picked. Since local errors are perturbations about the true time delay, they can be usefully characterized by a root-mean-square (RMS) error value. When a gross error occurs and an incorrect peak is chosen, however, that peak can occur at a location unrelated to the true peak and potentially very far from it. Because of this, we care primarily about whether or not a gross error has occurred, and not its magnitude, so a natural way to characterize gross error performance is by their frequency of occurrence.

Ianniello [48] suggested that the cutoff between local errors and gross errors should be on the order of the inverse signal bandwidth, so we choose a cutoff of 250 μs (1 / 4000 Hz), and call all errors with smaller magnitudes local errors and all errors with larger magnitudes gross errors. We report the RMS error for local errors and the observed frequency of occurrence for gross errors.

4.3.1 Synthetic data

Table 4.4 provides a concise summary of the relative performance of all of the time delay estimators on synthetic data. To generate this table, we normalized the error

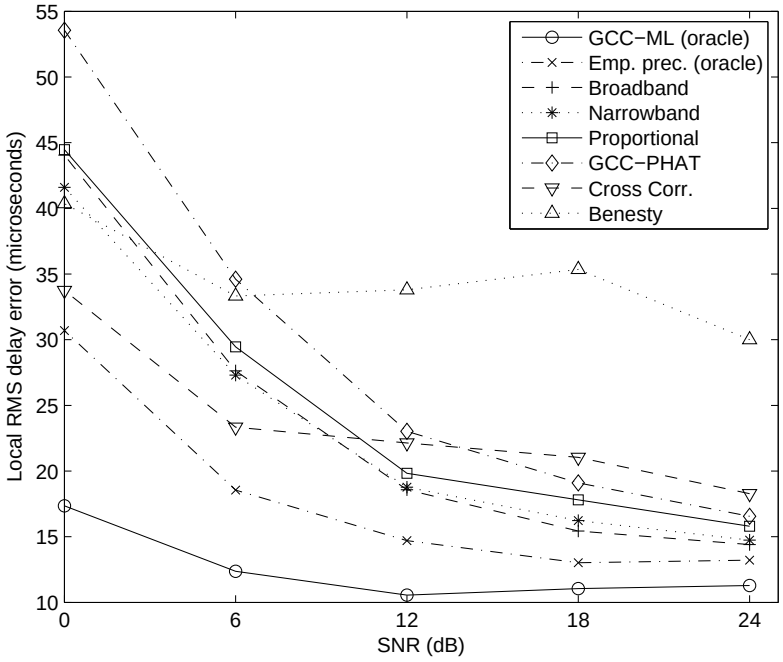
Technique	Norm. Local RMS Error	Norm. Gross Error Frequency
GCC-ML (oracle)	0.47	0.27
Emp. Prec. (oracle)	0.72	0.64
Broadband	0.86	0.81
Narrowband	0.86	0.79
Proportional	0.94	0.93
GCC-PHAT	1.00	1.00
Cross Correlation	1.47	1.49
Benesty	2.03	1.79

Table 4.4: Average normalized localization error in synthetic rooms. Error in in each room/noise level condition was divided by the GCC-PHAT error for that condition, and these normalized errors were then averaged across all conditions.

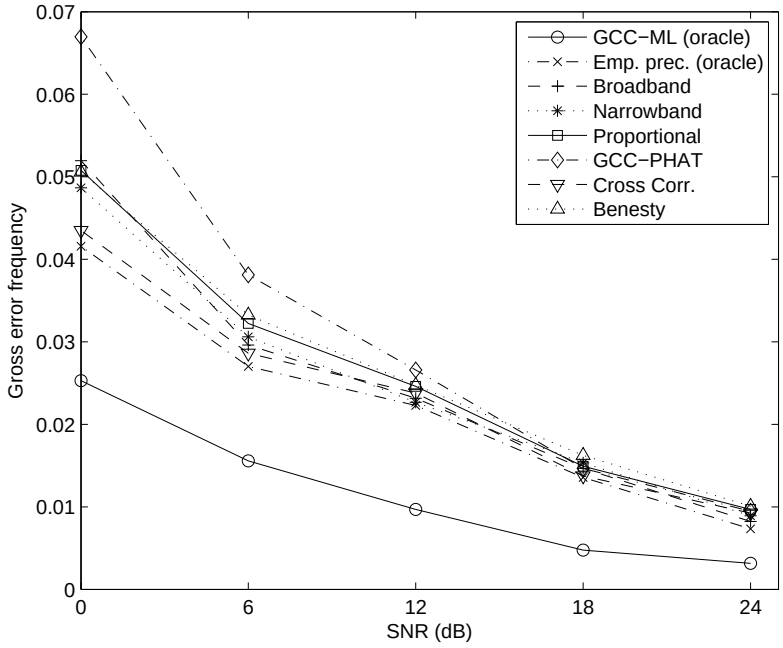
of each technique by the GCC-PHAT error for each specific room and noise level. We then average these normalized errors across all rooms and noise levels. By doing this normalization, we emphasize average relative performance. Without such normalization, small relative performance differences at low SNR would dominate large relative performance differences at high SNR.

Figures 4-2 through 4-6 show the performance of the time delay estimators across a range of reveberation times and background noise levels. Each figure shows the results for a single room, with room reverberation times increasing from Figure 4-2 to Figure 4-6. "SNR" on the horizontal axis refers to the level of the uncorrelated additive background noise.

The most pronounced trend is that error tends to decrease with increasing SNR for all techniques, although the decrease is not as pronounced for "Cross correlation" and "Benesty," the two worst-performing techniques. Unweighted cross correlation is known to perform poorly in reverberant environments [7], and since the amount of reverberation is only a function of the room, not the additive noise SNR, it is not surprising that unweighted cross correlation does not yield as much improvement as other techniques under high SNR conditions, where errors due to reverberation dominate. The only case in which unweighted cross correlation performs well is under low-reverberation, low-SNR conditions. Unweighted cross correlation emphasizes frequencies with more energy, and in the low-reverberation, white background noise case, frequencies with higher energy will have higher SNR and should therefore be

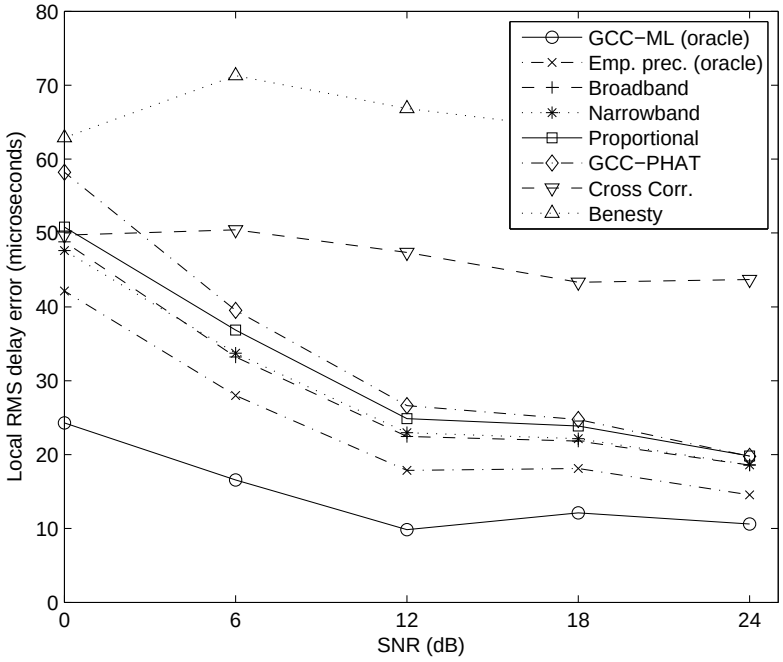


(a) Local RMS error

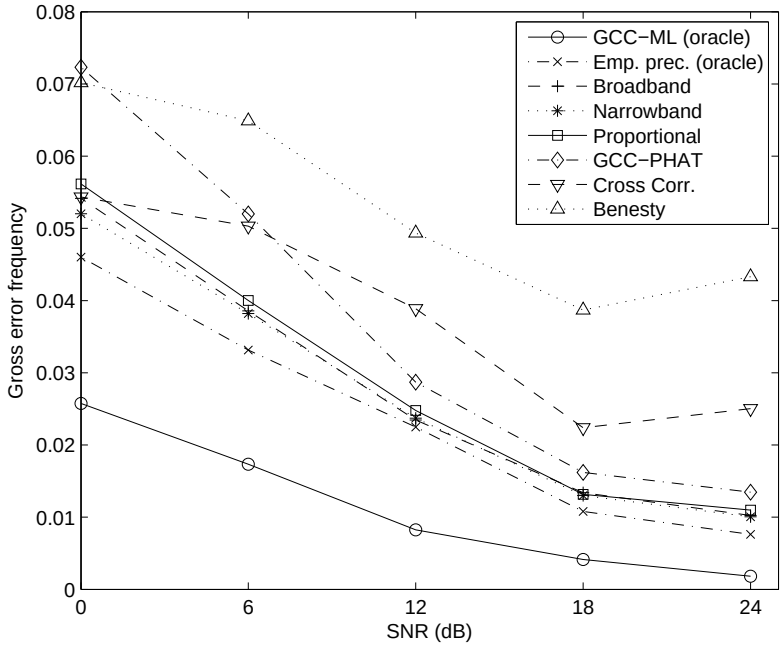


(b) Gross error frequency of occurrence

Figure 4-2: Localization performance in simulated room A, with an RT_{60} of 100 ms. "SNR" is the level of the additive white noise.

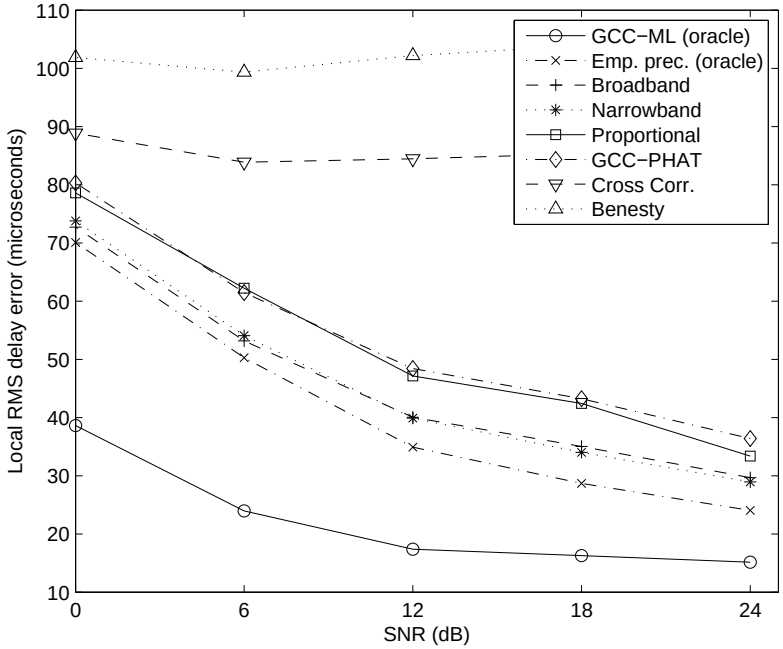


(a) Local RMS error

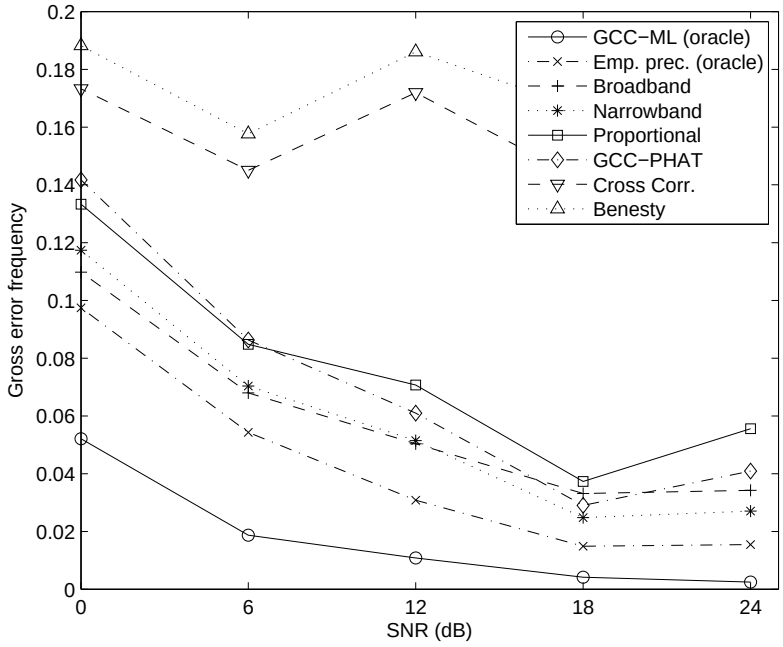


(b) Gross error frequency of occurrence

Figure 4-3: Localization performance in simulated room B, with an RT_{60} of 200 ms. "SNR" is the level of the additive white noise.

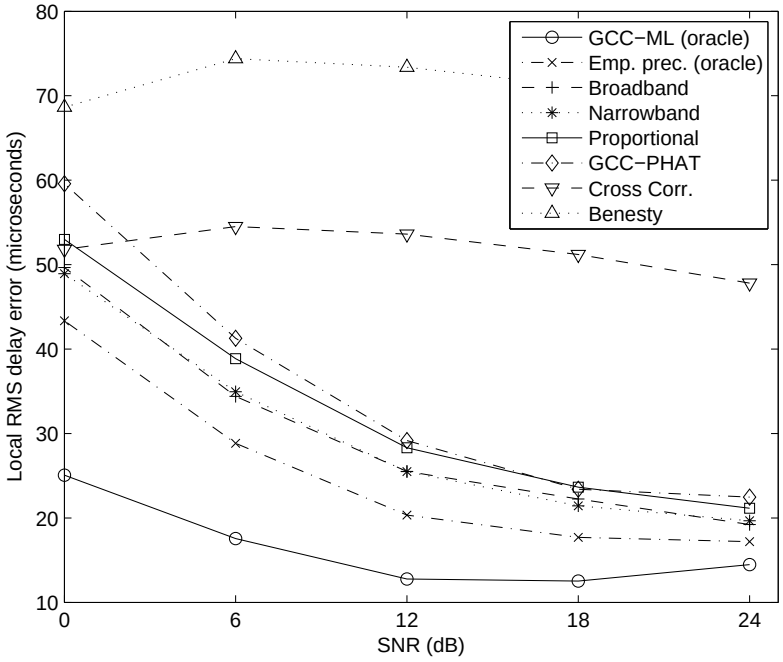


(a) Local RMS error

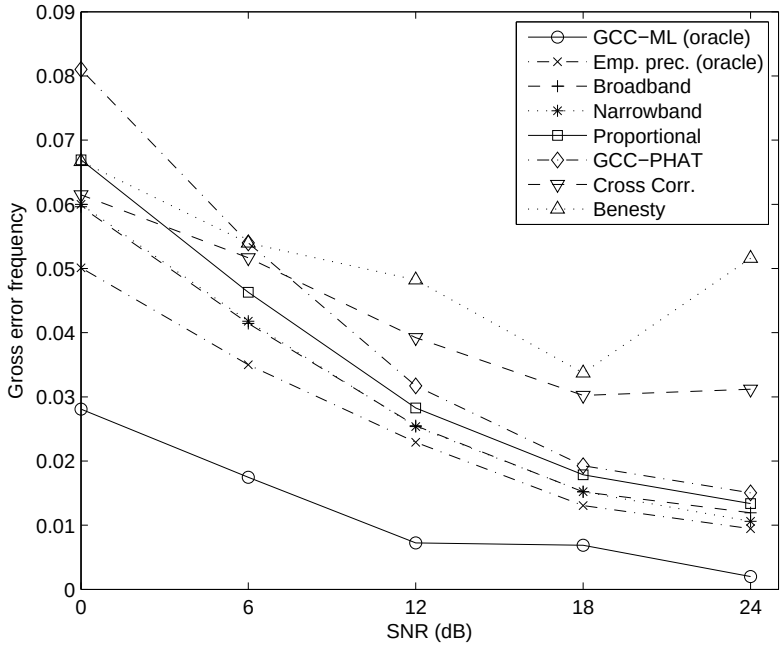


(b) Gross error frequency of occurrence

Figure 4-4: Localization performance in simulated room C, with an RT_{60} of 400 ms. "SNR" is the level of the additive white noise.

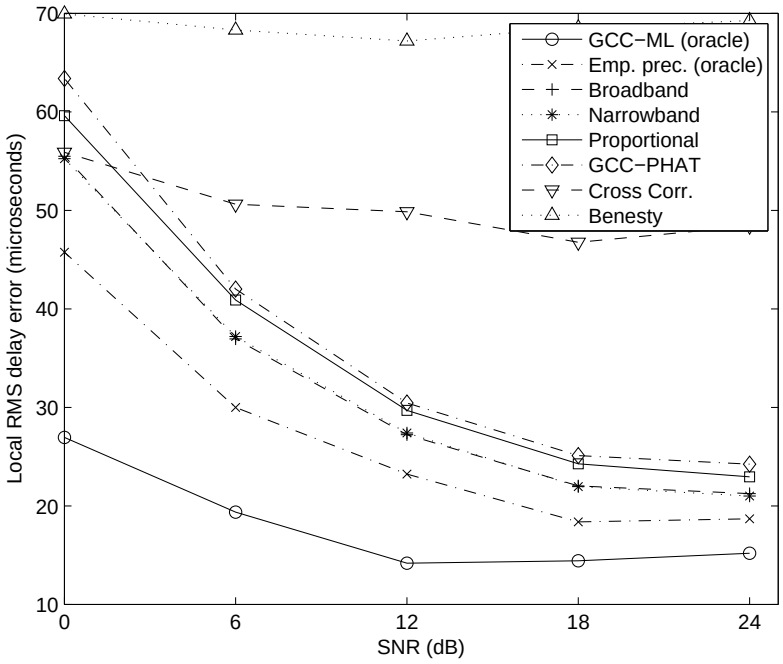


(a) Local RMS error

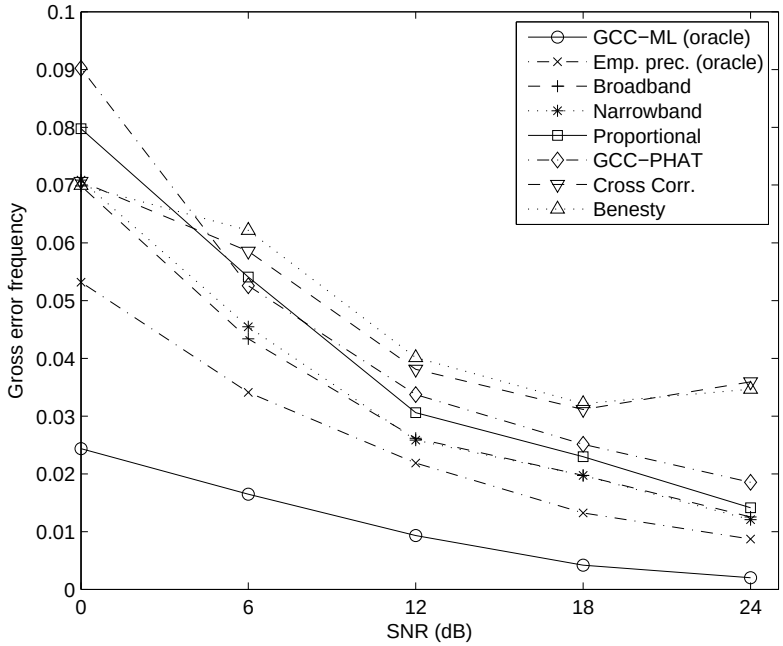


(b) Gross error frequency of occurrence

Figure 4-5: Localization performance in simulated room D, with an RT_{60} of 800 ms. "SNR" is the level of the additive white noise.



(a) Local RMS error



(b) Gross error frequency of occurrence

Figure 4-6: Localization performance in simulated room E, with an RT_{60} of 1600 ms. "SNR" is the level of the additive white noise.

emphasized.

In [7], Benesty demonstrated his technique's improved performance compared to GCC-PHAT in reverberant environments. In our experiments, however, we find that Benesty's technique performs worst overall. This is most likely because Benesty's experiments evaluated each algorithm's performance on 5-second-long audio segments, while our experiments evaluate performance on half-second-long audio segments. Because his technique is in some sense explicitly estimating the reverberant transfer functions of the two channels, it makes sense that it would require more data to converge compared to GCC techniques. Another potential drawback of Benesty's technique is that it requires that the signal covariance matrix be full rank. For the case of a harmonic signal, the signal covariance matrix will not be full rank, and approximately harmonic signals can have ill-conditioned covariance matrices. Longer audio segments are likely to have voiced and unvoiced speech as well as some variation in fundamental frequency, but short segments may be dominated by a single approximately harmonic vowel, causing problems for Benesty's technique.

Among the classic non-oracle-based GCC techniques, "GCC-PHAT" exhibits the best performance. This is consistent with previous findings on the performance of time delay estimators in reverberant environments [23]. Also consistent with previous findings is GCC-PHAT's poor performance in the low-reverberation, low-SNR case. This is exactly the opposite of the performance of the unweighted cross correlation since, in contrast to unweighted cross correlation, GCC-PHAT weights phase information at all frequencies equally regardless of signal energy. This is clearly the wrong thing to do in additive noise at low SNR, but it turns out to be the right thing to do for stationary signals in purely reverberant noise (as we showed in Section 2.2.2).

Human speech and many other auditory signals of interest are not stationary, however. This suggests that there may be room for improvement over GCC-PHAT in reverberant environments, and in fact we find that this is the case with our learned localization precision estimates. All three variants, "proportional," "narrowband," and "broadband," show marked improvement over GCC-PHAT.

With its single-tap filter, the proportional mapping can approximate both GCC-

PHAT and unweighted cross correlation as special cases. By setting all of its mapping coefficients to zero, its localization precision estimate $\hat{\sigma}^{-2}(f)$, and therefore its phase weighting, becomes a constant independent of the signal, which is exactly GCC-PHAT. Alternatively, by setting the filters at all frequencies to an appropriate constant, we can weight the phase information at each frequency proportional to the signal power. Unweighted cross correlation (unweighted in the sense that $\Psi(f) = 1$) effectively emphasizes phase information proportional to the intermicrophone cross-power, but for microphone pairs whose separation is small relative to the source distance, the cross-power and the single channel power will be similar. Since GCC-PHAT is a special case of the proportional weighting and since unweighted cross-correlation can be approximated as a special case, we should expect the proportional weighting to perform at least as well as those two standard techniques as long as the optimization criterion that we use is reasonable. Table 4.4 shows this improvement in average performance. As Figures 4-2 through 4-6 show, the proportional weighting's performance is generally similar to GCC-PHAT's, with most of its performance improvements coming at low SNR, where GCC-PHAT's performance is poorest.

Now we examine the "narrowband" and "broadband" mappings. These mappings have larger extent in time than the proportional mapping and can therefore capture relationships between localization precision and patterns of time-varying short-time signal spectra caused by speech nonstationarity.

For these experiments, the narrowband mapping had a temporal extent of 380 ms (a 101-tap filter on a spectrogram with a 3.75 ms frame step size). Larger temporal extent was not found to improve localization results in preliminary experiments. This is plausible since beyond the time scale of a few hundred milliseconds, lower level articulatory relationships diminish in importance and higher level linguistic relationships begin to dominate [88].

For a given temporal extent, a broadband filter will have N_f times as many taps as the narrowband filter, where N_f is the number of frequency bins in the spectrogram representation. Thus, for computational reasons, we were forced to limit the broad-

band filter extent in these experiments to 80 milliseconds. (Even if longer broadband filters were computationally feasible, because of their larger number of parameters, they would require more training data.)

As Table 4.4 shows, the overall performance of the narrowband and broadband mappings is essentially identical, and both show a significant improvement over the other non-oracle-based techniques. Figures 4-2 through 4-6 show that their performance is also very similar as a function of room reverberation and noise level. Since their performance is comparable, in practice this argues for the use of the narrowband mapping, which, even with a larger temporal extent, requires less computation because of its smaller extent in frequency.

Thus far we have demonstrated the superiority of our technique compared to other techniques. Now we empirically address the question of whether there is additional room for improvement if we had perfect knowledge of the statistics of the target signal and the noise. (We will see that it does.)

"GCC-ML" is the result of localizing with perfect knowledge of the short-time signal and noise magnitudes, but without knowledge of their phases. This information is unlikely to be available in practice, but the performance of GCC-ML shows us two things. First, it reminds us that even with perfect knowledge of the magnitude statistics, we cannot achieve perfect time delay estimation. (If we knew the exact magnitude and phase of the noise, we could subtract it out and achieve perfect localization performance.) Second, it shows that the GCC-ML weighting, which was derived for the case of Gaussian random signals in uncorrelated Gaussian noise and infinite window length, still works in practice for short-time windowed speech signals in a combination of reverberant and additive noise.

Finally, "Empirical precision" is the result of localizing based on empirical sample-based estimates of the localization precision. If these utterances had been in the training set, this would have been the target values that we regress to during training. Because the learned mappings do not perfectly predict the localization precision, we do not achieve the "empirical precision" performance in practice. The GCC-ML technique always has by far the best performance since it is signal-specific. The

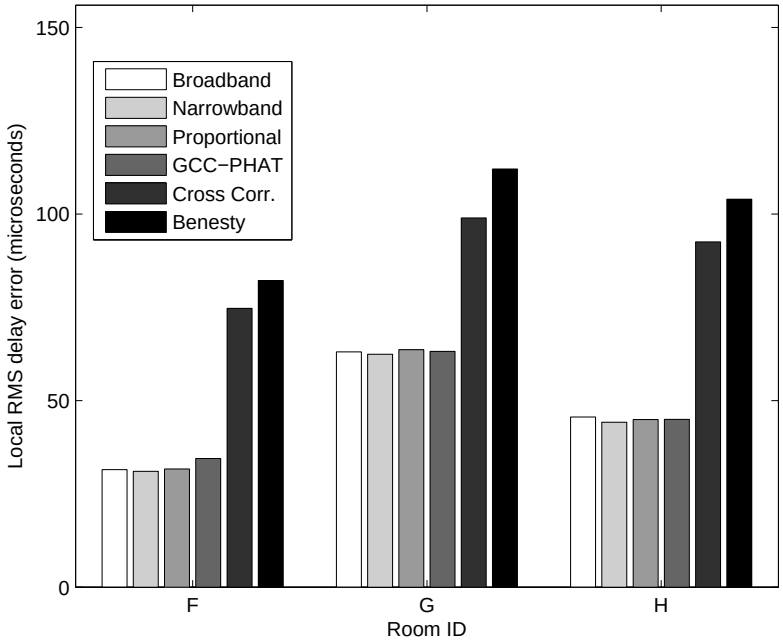
empirical precision technique underperforms it because it uses a precision estimate based on a sampled ensemble average.

To summarize, in all cases except for the combination of very low reverberation and very high noise, the narrowband and broadband mappings outperform all non-oracle-based techniques. In addition, the fact that our narrowband and broadband mappings outperform the proportional mapping shows that there is a practical benefit to using these richer mappings which are sensitive to energy distribution across time and/or frequency.

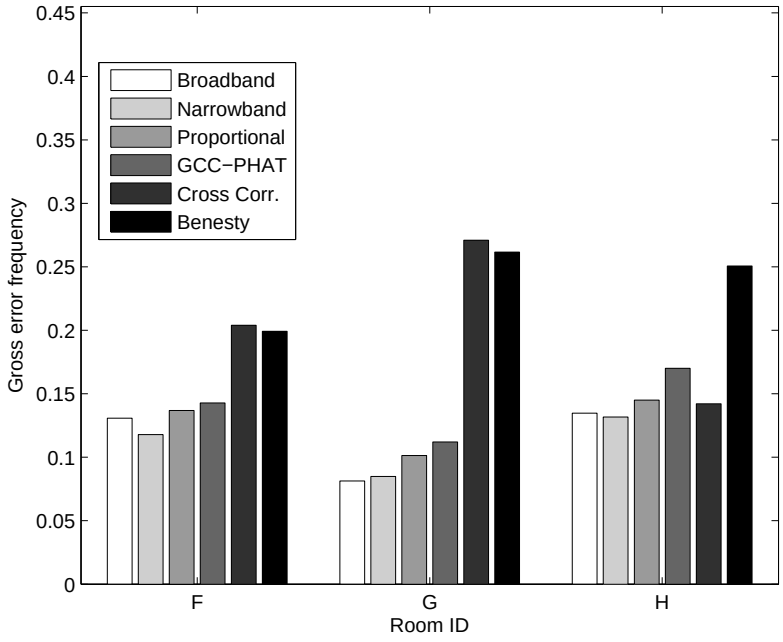
4.3.2 Real data

The results on synthetic data allowed us to explore the effects of varying noise and reverberation levels on time delay estimator performance. We now look at results on real data to convince synthetic data skeptics, of which we are one, that our technique really works. Table 4.5 summarizes results on real data in the same way that Table 4.4 summarized the synthetic results. Figure 4-7 shows separate results for the three rooms that we tested.

The real results, particularly the "Normalized Gross Error" in Table 4.5, are consistent with the synthetic results. It is again true that all variants of our technique outperform the standard techniques on average and that the broadband and narrowband mappings outperform the proportional mapping. The main discrepancy between real and synthetic is that for all of our mappings, the improvement in local RMS error is much smaller. This is likely due to the fact that in these experiments were done on real people who may have been slightly off from their assigned position during the data collection or who may have shifted during the recording process. The RMS time delay errors of a few tens of microseconds correspond to an angular error of a few degrees, so small source positioning errors could obscure some of the local RMS error performance differences. In spite of this, our proposed techniques still demonstrate some improvement.



(a) Local RMS error



(b) Gross error frequency of occurrence

Figure 4-7: Localization performance on data collected from the real rooms listed in Table 4.2.

Technique	Norm. Local RMS Error	Norm. Gross Error Frequency
Broadband	0.98	0.81
Narrowband	0.96	0.79
Proportional	0.98	0.91
GCC-PHAT	1.00	1.00
Cross Correlation	1.93	1.56
Benesty	2.16	1.74

Table 4.5: Average normalized localization error in real rooms.

4.4 Relationship to the precedence effect

Now that we have shown that our technique improves time delay estimation performance, we turn to the question of whether the resulting system bears any relation to the psychoacoustics of the precedence effect. We start by looking at the relationship between the reverberant speech spectrogram and the localization precision-gram. Figures 4-8 through 4-10 show example speech spectrograms and corresponding precision-grams for three of our five simulated rooms for the lowest-noise condition. In Figure 4-8, the 100 ms reverberation case, we see that the precision-gram is roughly proportional to the spectrogram. Wherever there is strong speech energy, there is precise localization information. In Figures 4-9 and 4-10, corresponding to 400 ms and 1600 ms reverberation times, respectively, note first that the overall precision values are lower because of the error introduced by reverberation. Next we see that, particularly in Figure 4-9, the precision-gram values are highest at the times of energy onsets in the spectrogram. This trend is less obvious in Figure 4-10 because room E, which is meant to model a large conference room, actually has a comparable amount of reverberant energy to room C, but room E's energy is spread out over a longer time.

To see whether our learned filters can capture this relationship, we next look at mappings learned in each of these reverberant conditions, as shown in Figure 4-11. Each subfigure shows narrowband mappings for a particular reverberant condition, and for each condition, mappings from a representative subset of the frequency bins are shown. The magnitudes of the filter coefficients for the 100 ms reverberation time case are much larger because, as can be seen in Figures 4-8(b) through 4-10(b),

the variance of the precision estimates in 100 ms reverberation case is much larger. (There are many high-precision time-frequency regions and many low-precision time-frequency regions in the 100 millisecond condition. There are mostly low-precision time-frequency regions with only a few high-precision regions in the 400 and 1600 millisecond cases since the reverberation makes the steady state sounds less useful for localization.)

In all cases the filter is approximately a superposition of a low-passed delta function and a band-passed edge-detector, as depicted schematically in Figure 4-12(b). The low-passed delta function component indicates that louder sounds provide better localization cues since for a mapping consisting solely of a delta function, a larger input (louder sound) will produce a proportionally larger output (higher-precision localization cue). This is to be expected in the presence of additive noise, where the ML weighting is proportional to the SNR and the SNR in our scenario is roughly proportional to the signal energy. The band-limited edge-detector can be interpreted as an onset detector, which is consistent with the precedence effect.

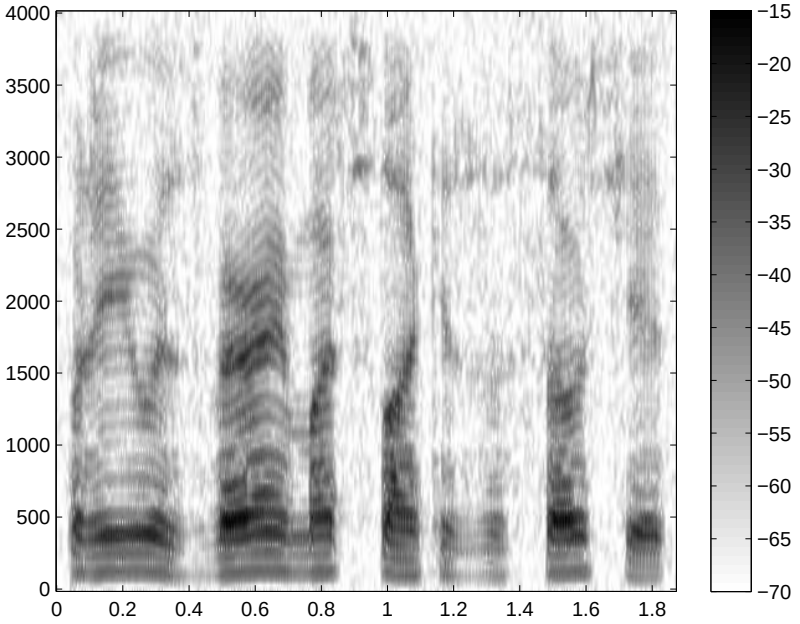
The relative amplitudes of the delta function and the edge detector reflect the relative importance of these two effects at each frequency and for each training condition's reverberation time. For the 100 millisecond training condition, the two higher frequency bins have almost no edge detector characteristic since this level of reverberation is relatively benign. The 400 millisecond condition shows some edge-detector characteristic at all frequencies because of the higher level of reverberation. Because of the longer time scale of the 1600 ms reverberation, the edge-detector characteristic is not as obvious around time 0, but instead takes the form of small negative values for almost all of the positive-time taps.

A representative subset of the actual filters used in our experiments is shown in Figure 4-12(a). These filters, which were trained on data from all reverberant conditions and noise levels, must take some shape that represents a compromise among all of the data collection conditions. As such, the edge-detector characteristic is subtler but still present.

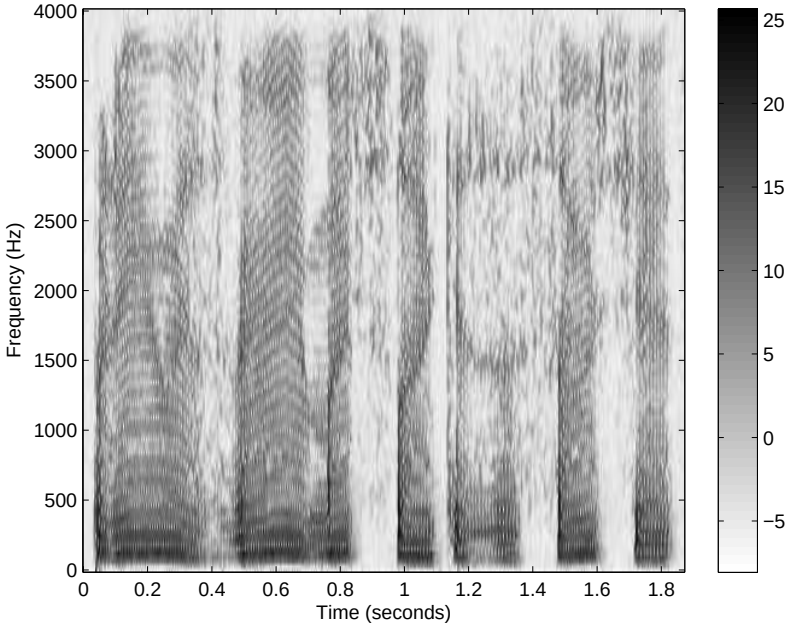
Our results are consistent with the precedence effect, and by looking at some

examples of the learned broadband mappings in Figure 4-13, we can see that they also can learn some (very simple) structure that is specific to the speech signal itself. First note that we again have some edge detector behavior in that we have mostly positive filter taps for $t \leq 0$ and mostly negative filter taps for $t > 0$. Next we see that since most of the filter energy is in taps corresponding to the target frequency bin, the predicted localization precision at a given frequency depends mostly on the spectrogram values at that frequency. This makes sense because the localization cues themselves are derived from the signal information at that frequency. An interesting but subtler effect is that there is some spread of filter energy across frequency. (Note the positive filter taps across frequency in all filters at $t = 0$.) Thus, even though there is no fundamental physical relationship between localization precisions at different frequencies, the correlation of signal energy across frequency in speech has been exploited in these broadband mappings. If energy fluctuations across frequency were not correlated, for example in the case of stationary signals, we would not see this structure in the broadband mappings. Cross-frequency dependencies in the precedence effect have been observed, for example in [84].

Finally, psychoacoustic research has found that (depending on the test criterion and stimuli used) the precedence effect acts to suppress post-onset localization cues for between 5 and 50 milliseconds [59]. In the narrowband filters shown in Figure 4-12(a), almost all of the filter energy (in both negative and positive taps) is within 50 milliseconds of time 0. Our system has implicitly learned the characterization of an "onset" that can provide precise localization over the range of acoustic environments on which we have trained, and its time scale is consistent with psychoacoustic findings.

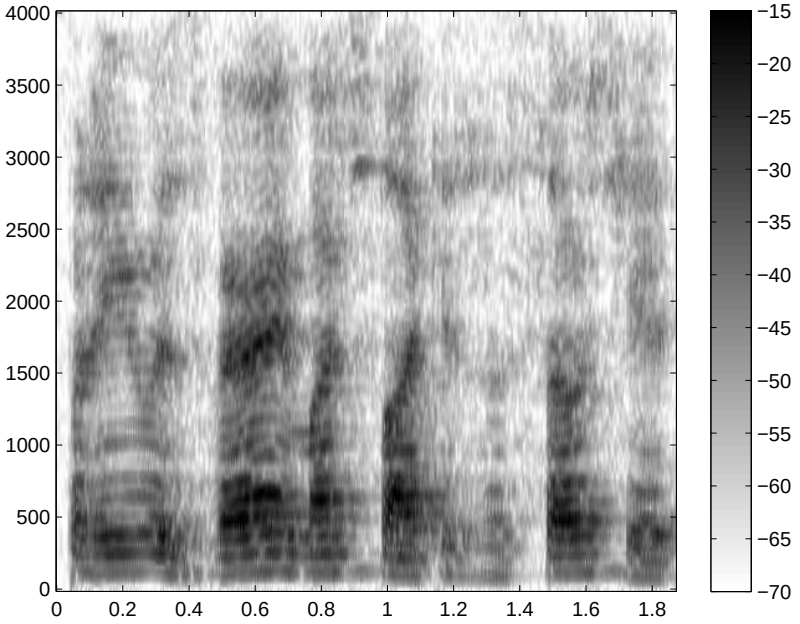


(a) Speech spectrogram

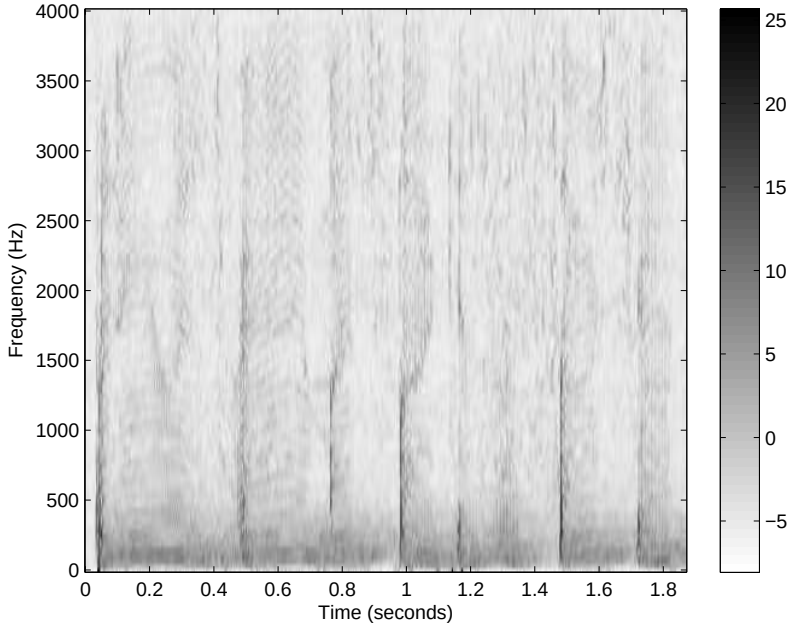


(b) Localization precision-gram

Figure 4-8: Sample speech spectrogram and corresponding localization precision-gram from simulated room A, with an RT_{60} of 100 ms. The male speaker is saying "So he was very much like his associates."

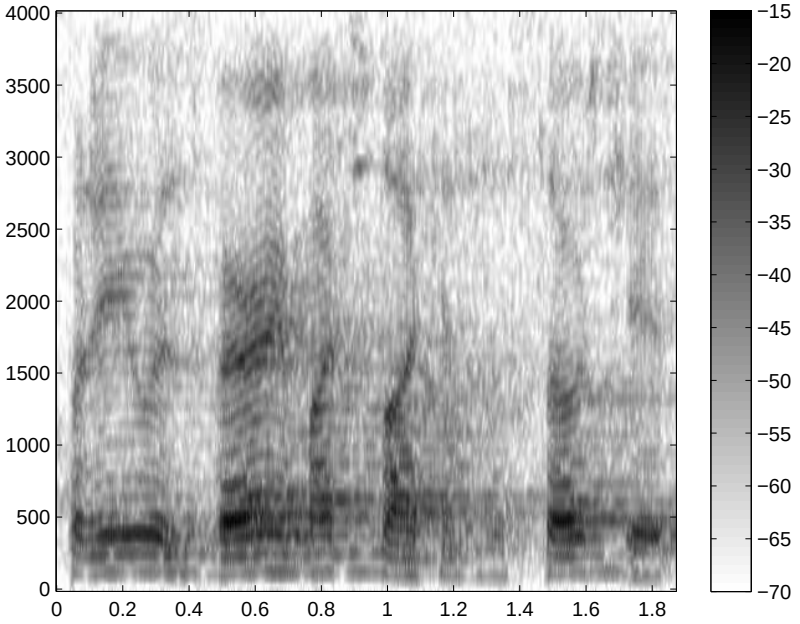


(a) Speech spectrogram

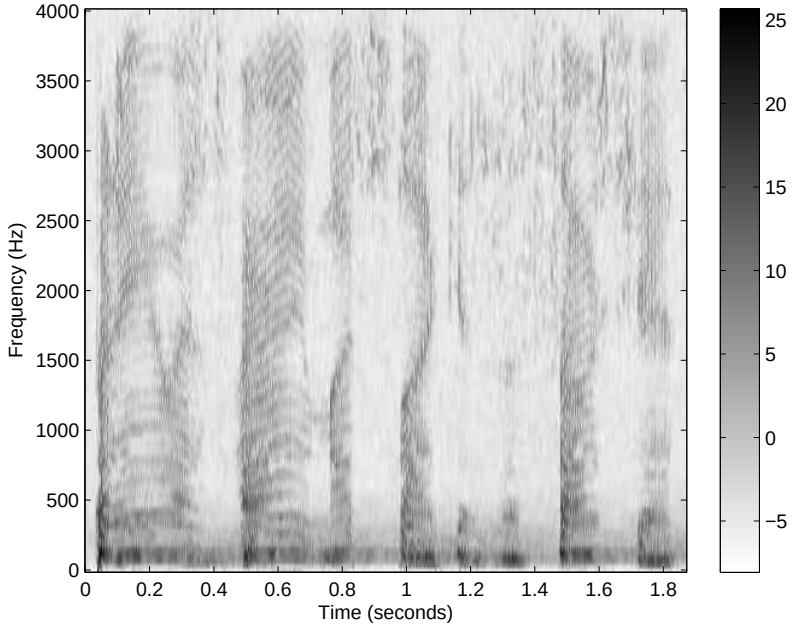


(b) Localization precision-gram

Figure 4-9: Sample speech spectrogram and corresponding localization precision-gram from simulated room C, with an RT_{60} of 400 ms. The male speaker is saying "So he was very much like his associates."

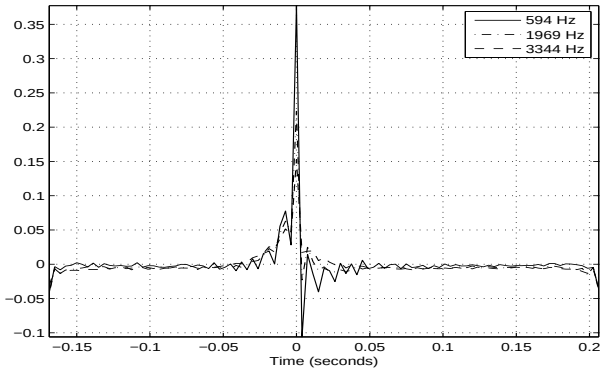


(a) Speech spectrogram

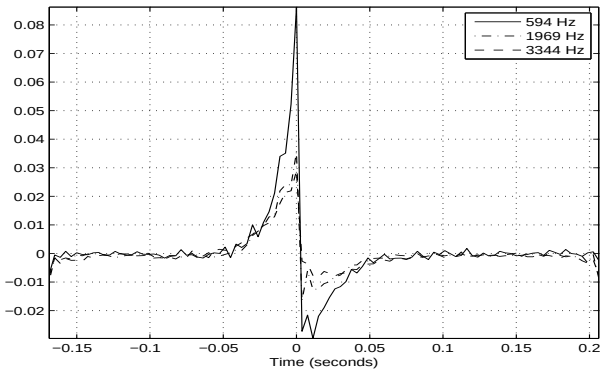


(b) Localization precision-gram

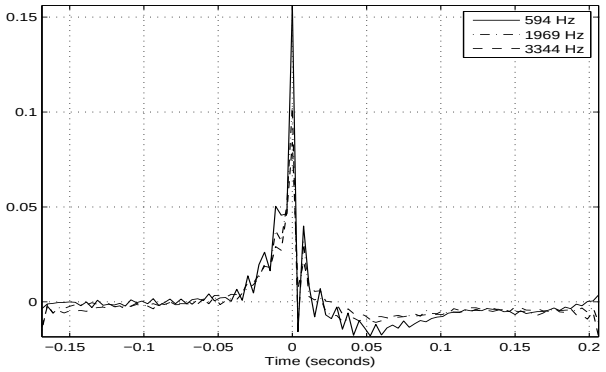
Figure 4-10: Sample speech spectrogram and corresponding localization precision-gram from simulated room E, with an RT_{60} of 1600 ms. The male speaker is saying "So he was very much like his associates."



(a) 100 ms reverberation time

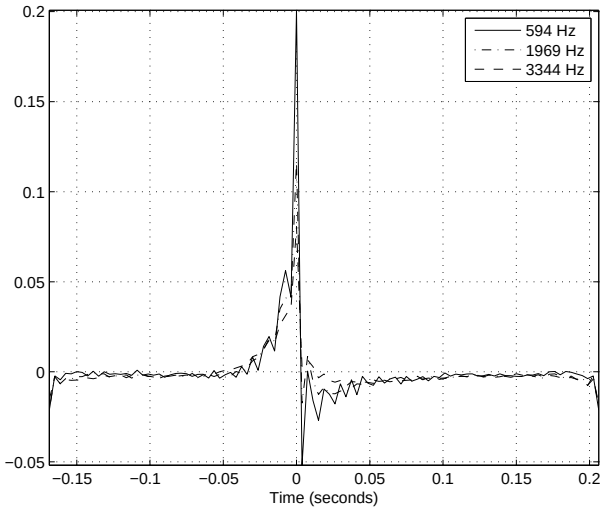


(b) 400 ms reverberation time

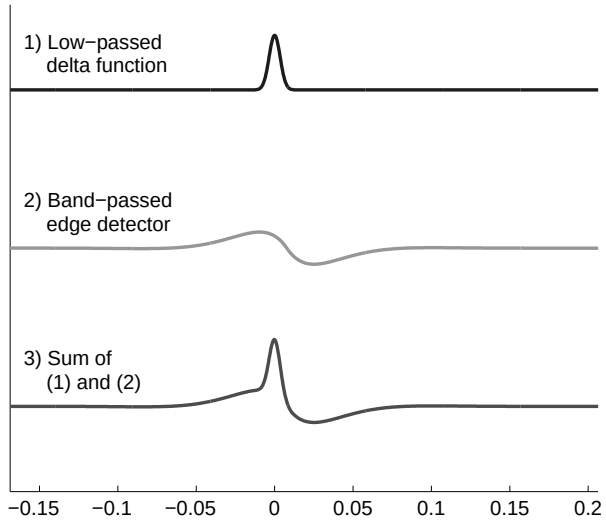


(c) 1600 ms reverberation time

Figure 4-11: A representative subset of narrowband filters for different reverberant conditions. Each subplot shows filters trained only on data from a single room. Within each subplot, three representative frequency bands are shown.

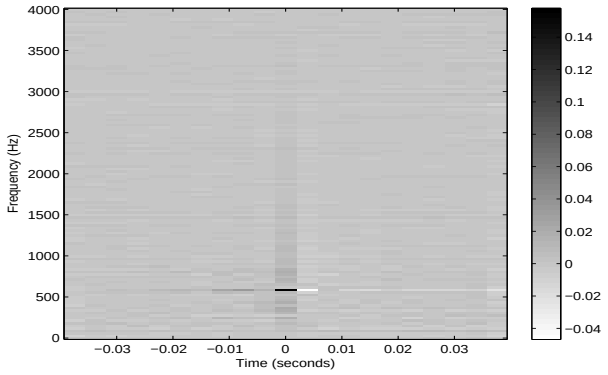


(a) Learned narrowband filters

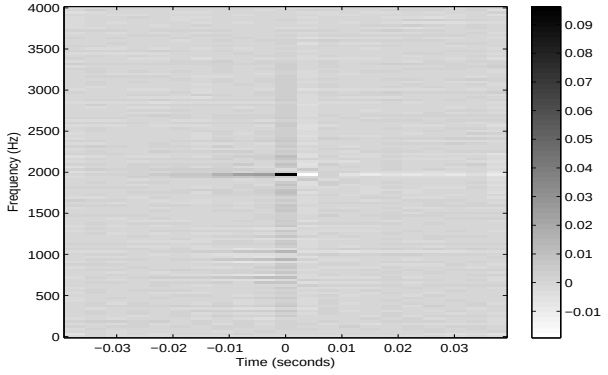


(b) Schematic decomposition

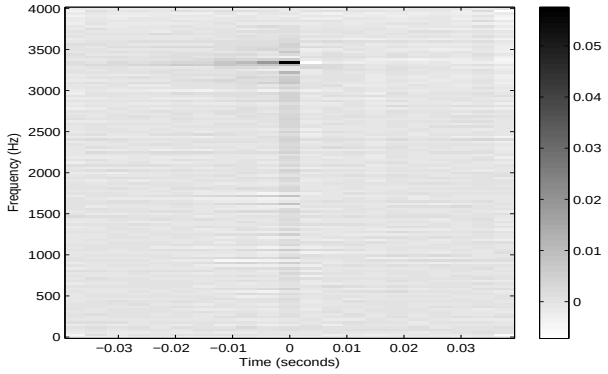
Figure 4-12: (a) shows the narrowband filters that result from training on all noise and reverberation conditions. (b) shows a schematic decomposition of the learned filters. Each of the learned narrowband filters can be viewed as a linear combination of a low-pass filtered impulse (top) with a band-pass filtered edge detector (middle). The bottom curve shows the linear combination of the top two curves, which is similar in shape to the learned narrowband filters.



(a) Freq. bin 20 (594 Hz)



(b) Freq. bin 64 (1969 Hz)



(c) Freq. bin 108 (3344 Hz)

Figure 4-13: Learned broadband filters for three representative filter bands. These filters have most of their energy in the frequency bin whose precision they are estimating, but there is some energy across all frequency bins, indicating that useful information is being integrated across frequency when calculating the optimal mapping.

Chapter 5

Source Separation

To this point, this dissertation has focused on improving the localization of a single source in a noisy, reverberant environment. This chapter demonstrates another use for our localization precision estimates by applying them to the simultaneous source separation problem.

5.1 The source separation problem

The isolation of a single auditory source from a mixture is one of the holy grails of audio processing. In what is known as the "cocktail-party effect," a term introduced by Cherry [19] and reviewed in [41], human listeners are able to focus on and comprehend one speaker in an environment filled with many simultaneous speakers. The human attentional mechanism allows a listener to focus on only one stream at a time [15], but a listener can shift attention from one speaker to another.

Computer audio processing systems have no inherent attention mechanism and are not limited to focusing on one speaker at a time. Thus, a related problem in computer audition is simultaneous source separation, in which a mixture of several simultaneous speakers is separated into its constituent speech streams.

Source separation is difficult because interesting sources signals like human speech are complicated and very nonstationary, and because even when the source location is known, the full source-microphone transfer function is typically unknown because

it depends on the details of the acoustic environment.

5.2 Components of the solution

Source separation is a difficult and often ill-posed problem. Cherry [19] lists several factors that could contribute to successful cocktail party performance, including differing localization cues among the speakers, knowledge of the temporal dynamics of speech, differences among speakers' voices, and speech-related visual cues. A successful source separation solution will likely need to exploit several of these factors.

This chapter examines a solution to the source separation problem that uses the localization precision estimates developed in Chapter 3 in combination with a statistical model of the spectral shapes of speech sounds and a simple implicit model of the temporal dynamics of speech.

5.3 Previous work

Source separation is typically formulated as an optimization problem consisting of an objective function that measures how well the speech is separated, a set of separating functions among which we hope to find one that can successfully separate the sources, and an optimization technique for searching the set of separating functions for one that optimizes the objective function.

5.3.1 Separating Functions

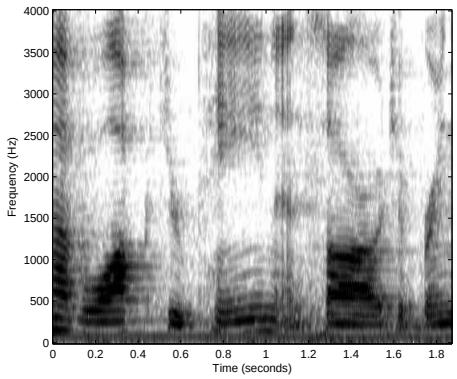
There are two common ways to separate simultaneous audio sources. The first method, which can completely separate sources only if there are at least as many microphones as sources, is to create a linear time-invariant multichannel unmixing filter which will imply a specific spatial filtering pattern. This method is used by beamformers [51] and most convolutive blind source separation (BSS) techniques (such as [71]).

The second method is to apply a time- and frequency-varying multiplicative mask to the spectrogram of the mixed input signal. Roweis [82] and Yilmaz and Rickard [93] use binary masks, in which each time-frequency region is either completely assigned to a given source or not assigned at all. Because of this, binary spectrogram masks depend on the sparsity of speech energy in the time-frequency domain and can only perfectly separate signals that are disjoint in this time-frequency representation. Empirical tests by both Roweis and Yilmaz and Rickard have demonstrated that oracle-chosen binary spectrogram masks can achieve excellent separation of two simultaneous speakers [82] and can achieve more than 9 dB improvements in SNR even for mixtures of ten speakers in an anechoic environment [93]. (Figure 5-1 shows an example of such an oracle-chosen binary mask.) Hershey and Casey [42] use a soft mask, in which each mask entry can take any value on the interval $[0, 1]$. This can be interpreted as a time-varying Wiener filter, and can in theory achieve better separation than the binary mask. In practice, however, binary masks work well because of the time-frequency sparsity of each individual speech signal.

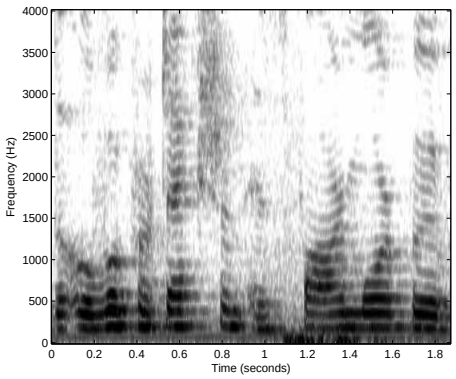
5.3.2 Objective Functions and Optimization Techniques

Objective functions for source separation may be defined independently for each separated source or as a joint function of all of the sources. Typically, a separating function and an objective function are paired because they permit the use of an effective optimization technique. We will describe the objective functions and optimization techniques together in this section.

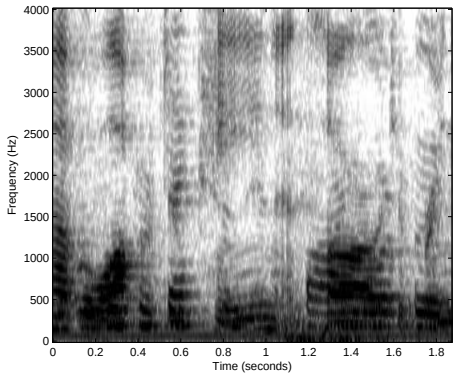
In the adaptive beamforming literature, the objective function is typically defined independently for each source and involves minimizing the output signal energy subject to array steering constraints [51]. If one knows the true direction to the target and can constrain the beamformer to always pass signals from that direction, one can minimize the noise level in the output signal by minimizing the total output signal energy subject to the steering constraint. In uncorrelated noise and with perfect knowledge of the source-to-array transfer function, this will act to minimize the noise present in the output signal. In many practical situations, however, the target direc-



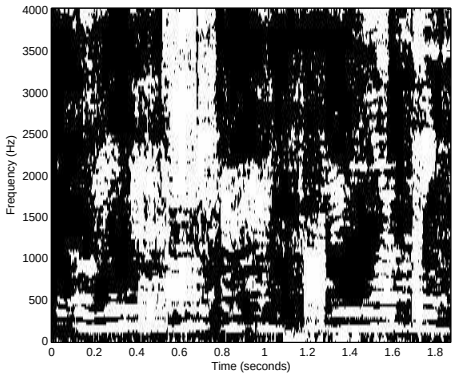
(a) Speaker 1 spectrogram



(b) Speaker 2 spectrogram



(c) Simultaneous speech spectrogram



(d) Ideal binary spectrogram mask

Figure 5-1: Example of a binary spectrogram mask. (a) and (b) show spectrograms for two isolated male speakers. Speaker (a) is saying "...have walked through pain and sorrow..." and speaker (b) is saying "...overprotection is far more..." (c) shows the spectrogram when speakers 1 and 2 speak simultaneously. Darker colors indicate higher energy. (d) shows the ideal binary mask for separating the two speakers. Black regions indicate where speaker (a) is louder and white regions show where speaker (b) is louder. The spectrogram in (c) can be multiplied by the mask in (d) and its binary complement to reconstruct the two individual speakers.

tion is not precisely known and the noise is not completely uncorrelated, so adaptive beamformer performance degrades significantly [14]. Additional constraints can be imposed to limit the sensitivity of adaptive beamformers to modeling errors [44], but these techniques achieve this robustness at the cost of reduced performance in the best case. There is a simple closed-form solution to this basic adaptive beamforming objective function [51] and many of its variants, although they are usually formulated as online adaptive filtering problems to allow for source motion and environmental changes.

In the blind source separation literature, the objective function is typically related to some measure of statistical independence among the reconstructed sources [24]. This simple assumption is both the blessing and the curse of these techniques. When we have only very limited knowledge of the signal statistics and mixing parameters, for example when analyzing poorly understood medical signals as in [9,72], this can be a practical objective function. However, when the mixing system has many parameters, as is the case with the many taps of the long filters associated with a reverberant environment, these objective functions and their weak assumptions about the sources may require a large amount of data to provide a useful measure of separation. In practice, the most successful current algorithms for blind speech source separation in reverberant environments employ gradient-based optimizations that converge reasonably quickly to local optima [16, 71].

To separate simultaneous sources, Yilmaz and Rickard propose the DUET source separation technique [93] which uses a per-spectrogram-bin objective function that assigns each bin to the source with which its localization cues are most consistent. This objective function is equivalent to a binary hypothesis test that can be easily solved given their use of histogram-based density estimates.

A number of systems have used objective functions related to the "speech-ness" of the separated signals. Ephraim [28] used hidden Markov model (HMM) speech models for speech enhancement. Roweis [81, 82], Hershey and Casey [42], and Reyes-Gomez et al. [77] train HMMs on isolated speech and then simultaneously decode multiple HMMs to separate multiple speech streams. [28], [82], and [42] use the forward-

backward algorithm to find marginal state distributions and then use these state distributions to separate or enhance the signals using a time-frequency mask. [77] uses the forward-backward algorithm as a subroutine in a procedure that optimizes a beamformer to separate multiple sources. Brandstein and Griebel [14, 37] use a multichannel extension of the Dual Excitation speech model [39] and a linear predictive model to exploit the harmonicity and formant structure of speech within a speech enhancement system.

5.4 Our Technique: Combining Localization Cues with Speech Models

Our goal in this chapter is to combine localization cues with a model of speech spectra and a simple model of speech temporal dynamics to facilitate source separation using a spectrogram mask. We will use a binary spectrogram mask to separate speakers, and our objective function will combine localization cue likelihoods with spectral shape likelihoods. This can be viewed as adding localization cues to the single-channel separation techniques of Roweis [82] and Hershey and Casey [42] or as adding a speech model to the DUET technique [93]. The key ingredient that allows for this combination is an estimate of localization cue errors across time and frequency, which was precisely what was developed earlier in Chapter 3.

Our technique is conceptually similar to the technique described in [67], which also combines localization cues with a state-space model of speech spectra. We differ from them in that they choose a particle filter implementation that takes 32 CPU days to process one second of audio (2.7 million $\times$ real-time). Our implementation uses simple temporal smoothing of localization cues and runs at less than 100 $\times$ real-time. We also evaluate our technique in more reverberant environments.

We will first describe each of the components of our solution. Then we will conclude this section with an overall summary of our algorithm.

5.4.1 Localization cues

The localization precision, $\sigma_{loc}^{-2}(u,f)$ (where u is a time index and f is a frequency index) discussed in Chapter 3 implies a simple generative model for microphone pair cross-power spectrum phase measurements once the source position is known. Specifically, assuming that the phase associated with the true source position is $\theta_{true}(u,f)$ and the phase noise is Gaussian, then the observed phase, $\theta(u,f)$, is distributed as

$$p(\theta(u,f);\theta_{true}(u,f)) = \mathcal{N}(\theta(u,f);\theta_{true}(u,f),\sigma_{loc}^2(u,f)) \tag{5.1}$$

One problem with this model is that phase noise cannot be Gaussian since phase is periodic with period 2π. Another problem is that we do not actually know $\sigma_{loc}^2(u,f)$; in practice we will use $\hat{\sigma}_{loc}^{-2}(u,f)$, the estimated precision resulting from applying the mappings developed in Chapter 3. We will see from our results that even after making these assumptions, we are able to separate speech reasonably well.

In a situation with two simultaneous speakers with true (direct path) phases θ_{true_1} and θ_{true_2}, one can define a binary mask, $M(u,f)$, to separate out source 1 by setting the mask to one everywhere that source 1 maximizes the likelihood of the observed phase and zero elsewhere. Source 2 can be separated out by applying the complementary mask, $(1 - M(u,f))$.

$$M_{DUET}(u,f) = \begin{cases} 1 & \text{for } (u,f) \text{ such that } p(\theta(u,f);\theta_{true_1}(u,f)) > p(\theta(u,f);\theta_{true_2}(u,f)) \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

In our system, the estimated $\hat{\sigma}_{loc}^2(u,f)$ is derived directly from the signal without knowledge of who is speaking. As a result, we have only a single estimate and must share $\hat{\sigma}_{loc}^2(u,f)$ across all sources. This may not always be realistic, for example when one speaker is much closer to the array and generates much more reliable localization cues. However, without knowledge of source distances or other specific acoustic conditions of the sources, using the same shared $\hat{\sigma}_{loc}^2(u,f)$, which comes from map-

pings trained on a variety of acoustic conditions and therefore represents an "average" value, seems reasonable. Because we are using a shared variance across all sources, picking the maximum likelihood source corresponds to picking the source whose true (direct path) phase is closest to the observed phase. Picking the maximum likelihood source is the DUET separating technique, although our Gaussian phase distribution implies a different likelihood function than specified in [93].

Depending on localization cues to independently specify the mask value at each time and frequency, as is done in DUET, works well in near anechoic environments, but Yilmaz and Rickard report that it does not perform well in strong reverberation. To achieve good performance in reverberant environments we must integrate cues across time and frequency.

5.4.2 Speech spectral model

To integrate information across frequency, we use a Gaussian mixture model of the speech log spectrum in each frame. We train a speaker-independent diagonal-covariance mixture model using expectation-maximization (EM) [21]. Under each mixture component, the log-spectrum is assumed to be distributed as

$$L_{m_{shape}}(u) = \prod_f \mathcal{N}(s(u,f); \mu_{m_{shape}}(f), \sigma^2_{m_{shape}}(f)) \tag{5.3}$$

where $s(u,f)$ is the log spectrum at time frame u, $\mu_{m_{shape}}(f)$ and $\sigma^2_{m_{shape}}(f)$ are the mean and variance associated with mixture component m. We use the *shape* subscript to indicate that these are the distribution parameters for the spectral shape, in contrast to the *loc* parameters for the localization cue distributions. The dotted and dashed lines in Figure 5-2 show examples of mixture component means from our model.

Once we have this isolated speaker mixture model, we use it to create a two-speaker Cartesian product mixture model that we can subsequently use for source separation. Given two isolated-speaker mixture models, we create a new two-speaker mixture model with one state for each possible pair of states from the isolated-speaker

models (Figure 5-2). A problem that arises in this combination is that our observation distributions are specified in the log-spectral domain, but uncorrelated sources add in the power domain. To derive an observation for the summed factorial state from the single-speaker distributions, Roweis [82] uses the "log-max" approximation, in which he assumes that $\log(s_1 + s_2) \approx \max(\log s_1, \log s_2)$. Hershey and Casey [42] exponentiate the log spectra and use moment matching to find a Gaussian distribution for the sum's spectral distribution. We follow the approach of [42], but in addition to deriving a two-speaker spectrum distribution, we also keep track of which speaker has the maximum energy at each frequency, as shown in Figure 5-2(b).

We decompose the likelihood for a given two-speaker mixture state into two terms, one for spectral shape and another for localization cues. For state m, the spectral shape term, $L_{m_{shape}}(u)$, is just the evaluation of the observed log spectrum under the distribution specified by the mixture state mean and covariance. The localization cue likelihood, $L_{m_{loc}}$, assumes that at each frequency, the speaker with a higher mean energy at that frequency will be responsible for generating the localization cue, so the observed phase is evaluated according to that speaker's model. The two likelihood terms are assumed independent given the state:

$$L_{m_{loc}}(u) = \prod_f p(\theta(u,f); \theta_{true_{mask(f)}}(u,f)) \tag{5.4}$$

$$L_{m_{tot}}(u) = L^{\alpha}_{m_{shape}}(u) L^{\beta}_{m_{loc}}(u) \tag{5.5}$$

where θ is the observed phase difference, $mask(f)$ is the dominant speaker mask shown in Figure 5-2(b), $\theta_{true_{mask(f)}}$ is the true phase associated with the dominant speaker in that frequency, and $L_{m_{tot}}$ is the overall likelihood of the data under that state distribution. α and β are likelihood weighting parameters whose values were chosen based on experiments with a validation data set. We evaluate the likelihood for all mixture components to find a posterior marginal distribution over the states.

Now we must use these likelihoods to create a separating mask. Hershey and Casey use the marginal state distributions to determine a marginal spectral distribution, and from that they compute a Wiener filter. Roweis finds the MAP state assignments and

uses these to generate a binary mask. Binary masks appear to work well; however, we have found binary masks based on the posterior marginals to achieve better separation than those based on the single MAP state. We define our binary mask to be

$$M_{GMM+loc.}(u,f) = \begin{cases} 1 & \text{for } (u,f) \text{ such that } \sum_m (\tilde{L}_{m_{tot}}(u) * mask_m(f)) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \quad (5.6)$$

$$\tilde{L}_{m_{tot}}(u) = \frac{L_{m_{tot}}(u)}{\sum_{m'} L_{m'_{tot}}(u)} \quad (5.7)$$

where Equation 5.7 normalizes $\tilde{L}_{m_{tot}}(u)$ to sum to 1.

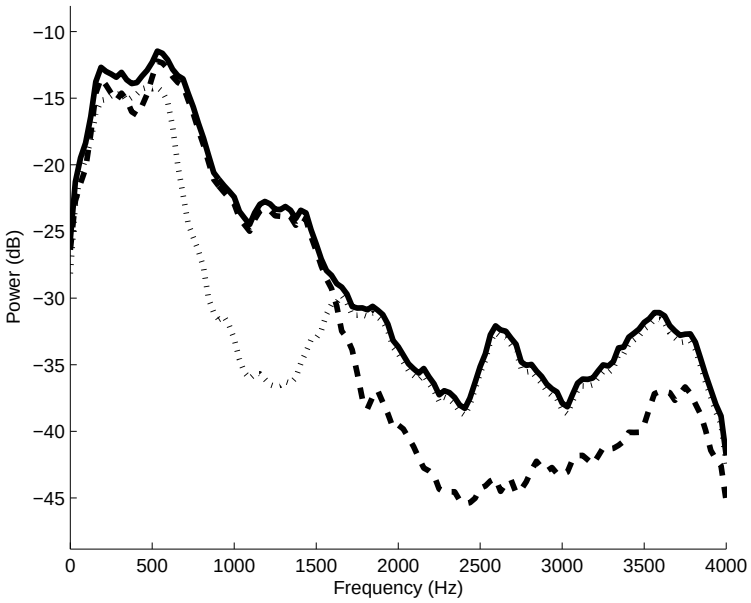
5.4.3 Temporal smoothing

In [82], Roweis trained HMMs on clean speech and used them to create a factorial HMM to separate simultaneous speakers. However, in [81] he subsequently stated that a Gaussian mixture model of speech spectra (equivalent to an HMM without dynamics) worked just as well as the HMM.

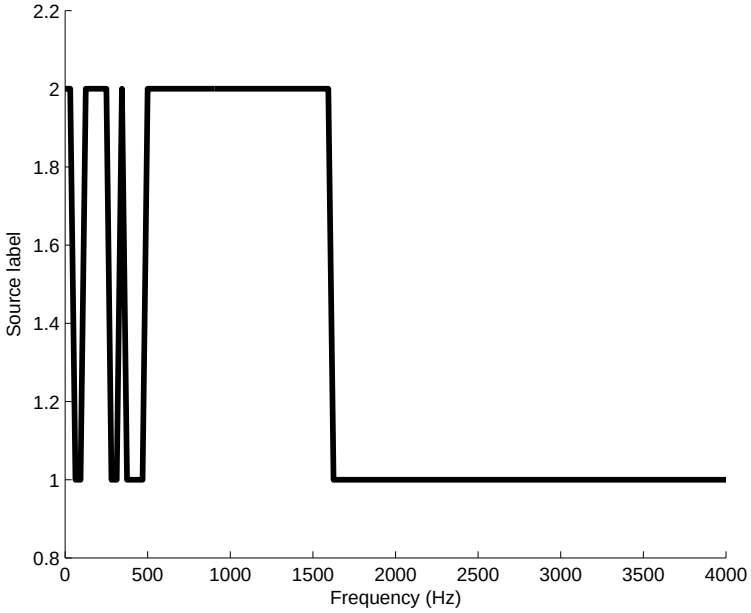
This suggests that the relatively short-term Markov state dynamics are not very useful for the source separation problem. Clearly, however, within a given frequency bin, the ground truth mask in Figure 5-1(d) often maintains the same value for many frames in a row, and this slow change implies longer time-scale dynamics. One way to capture these dynamics is to augment the speech model to use N^{th} order Markov dynamics, but for any reasonable N this would lead to an impractically large amount of computation.

A simple way to effectively capture some of these longer time-scale dynamics is to use an exponential forgetting factor to low-pass filter the location log-likelihood terms over time.

We define a smoothed location likelihood term as



(a) Spectral log magnitudes



(b) Max power source indices

Figure 5-2: Creation of a two-speaker mixture model component. (a) The two-speaker log spectrum (solid line) is formed from the power sum of the isolated speaker spectra (dashed and dotted). (b) The mask indicates which speaker is dominant at each frequency and is used both for separation and to evaluate localization cue likelihood.

$$p_{smooth}(\theta(u,f);\theta_{true}(u,f)) = p_{smooth}(\theta(u-1,f);\theta_{true}(u-1,f))^{(1-\gamma)} p(\theta(u,f);\theta_{true}(u,f))^{\gamma} \tag{5.8}$$

where γ is the forgetting factor. (In practice, we implement this as an autoregressive filter on log-likelihoods.) We use this to redefine our original location likelihood from Equation 5.4 as

$$L_{m_{loc}}(u) = \prod_f p_{smooth}(\theta(u,f);\theta_{true}(u,f)) \tag{5.9}$$

This is a form of likelihood weighting, which we also used in Equation 5.5. With our Gaussian distributions, this is equivalent to a multiplicative scaling of the variance. Therefore, older, more out-of-date observations will be incorporated with larger variances (tending toward a uniform distribution), and will have less effect on the overall likelihood.

5.4.4 Independence assumptions

We have made a number of independence assumptions that are clearly not true:

- We have assumed that spectrogram frames are independent even though they come from overlapping windows of the original signal and even though speech has significant temporal structure.
- We have assumed that frequency bins are independent even though the windowing used to generate them introduces cross-frequency dependencies and even though speech has broadband structure.
- We have assumed that the likelihood terms associated with spectral shape are independent of the likelihood terms associated with localization cues after conditioning on the state.

These independence assumptions lead to overcounting of evidence. A simple and often effective way to deal with this is to raise each likelihood term to some power

(or equivalently to multiply log likelihoods by a constant) to compensate for this overcounting. For example, this must often be done to effectively combine acoustic models and language models in automatic speech recognition [46, p. 610] [69]. We have incorporated likelihood weighting through the α and β parameters in Equation 5.5 and the forgetting factor, γ, in Equation 5.8. We have found log likelihood scaling to work well for our technique.

5.4.5 Algorithm summary

Here we summarize our algorithm. Prior to running the algorithm, it is assumed that we have generated $\hat{\sigma}_{loc}^{-2}(u, f)$ as described in Chapter 3 and have learned a Gaussian mixture model of speech spectra as described in Section 5.4.2.

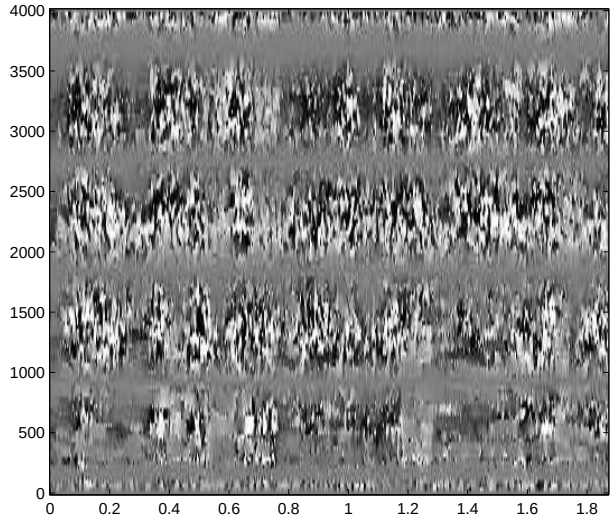
Algorithm summary:

1. Compute localization cue likelihoods (Equation 5.1) and smooth them (Equation 5.8). For illustrative purposes, these quantities have been used to compute log-likelihood ratios in Figure 5-3 for the example speech segment from Figure 5-1.

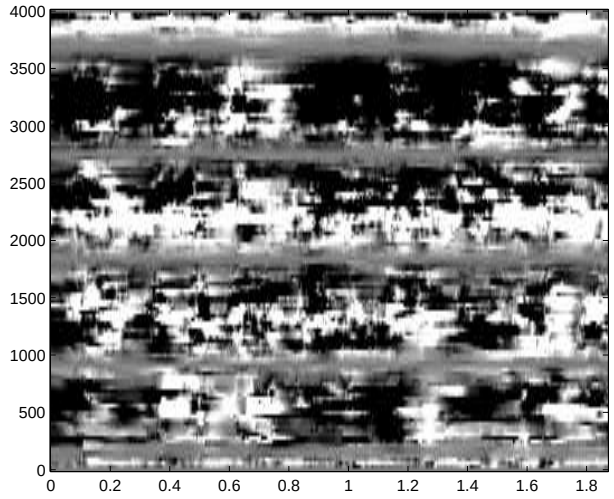
2. For each spectrogram frame, u, and for each GMM component, m, compute $L_{m_{shape}}(u)$ (Equation 5.3) and $L_{m_{loc}}(u)$ (Equation 5.9). Combine these to find $L_{m_{tot}}(u)$ (Equation 5.5).

3. Use these overall component likelihoods to compute a spectrogram mask $M_{GMM+loc.}(u, f)$ (Equation 5.6).

4. Pointwise multiply the spectrogram, $s(u, f)$, by the mask, $M_{GMM+loc.}(u, f)$ to separate out speaker 1. Pointwise multiply by the complementary mask, $(1 - M_{GMM+loc.}(u, f))$, to separate out speaker 2.



(a) Localization cue log-likelihood ratios



(b) Smoothed log-likelihood ratios

Figure 5-3: Localization cue log likelihood ratios for the example speech in Figure 5-1. (a) shows raw (unsmoothed) phase log-likelihood ratios that result from evaluating Equation 5.1 for each of the two sources and taking the logarithm of their ratio. (b) shows the smoothed phase log-likelihood ratios that result from evaluating Equation 5.8 for each of the two sources and taking the logarithm of their ratio. Lighter regions are where speaker 1 is more likely, and darker regions are where speaker 2 is more likely. Note that there is some rough correspondence between light regions in these figures and white regions in Figure 5-1(d).

5.5 Experiments

Source separation experiments were performed on both real and synthetic data. The goal of the experiments on synthetic data is to test the technique systematically over a range of acoustic environments and speakers. The experiments on real data show that the technique does not depend on any unrealistic assumptions of the simulation. The real data experiments cover a range of acoustic environments, but not as systematically as the results on synthetic data.

5.5.1 Experimental setup

The experiments were carried out in the same rooms used for the localization results in Chapter 4, and again all results are for audio sampled at 8 kilohertz. The speaker time-delay separations tested on synthetic data were 0.15 ms, 0.3 ms, 0.65 ms, and 1.1 ms, corresponding to broadside angular separations of 8°, 16°, 35°, and 61°, respectively. The 24 speakers in the TIMIT core test set were randomly combined into 12 pairs, and these 12 pairs were tested at all four angular separations in all five synthetic rooms. Roughly 30 seconds of audio was used in each configuration. For the real data, one pair of speakers was tested in each room, and each of these pairs was tested in 12 different configurations with separations ranging from 0.3 ms to 1.1 ms, or from 16° to 61°. Again roughly 30 seconds of audio was used in each configuration. All speakers in the real data experiments were male.

We compare seven techniques on this source separation task. The first technique, "Wiener filter," is an oracle-based technique in which, using knowledge of the pre-combined signals, we compute a time-varying Wiener filter and apply it to the mixed signal. The Wiener filter minimizes reconstruction error for stationary signals, so if our signals are approximately stationary for the time-scales of interest, the Wiener filter should yield the best SNR. In our Wiener filter implementation, we assume that the two speech signals are uncorrelated and compute a continuous mask (as opposed to other techniques' binary masks) as

$$M_{Wiener}(u,f) = \frac{|s_1^2(u,f)|}{|s_1^2(u,f)| + |s_2^2(u,f)|} \tag{5.10}$$

where s_1 and s_2 are the spectra for the two individual speakers.

The second technique, "ideal mask," is an oracle-based technique in which, using knowledge of the pre-combined signals, a binary mask is created for which each time-frequency region is assigned to the source that has the most energy at that frequency. This maximizes the SNR within each time-frequency region, and as Yilmaz and Rickard [93] and Roweis [82] report, its subjective performance is also quite good. This technique is equivalent to a binary thresholded Wiener filter.

"GMM + loc." is our technique as described earlier in this chapter. For these experiment, we used a 40-component Gaussian mixture model, which was computationally reasonable and yielded good results.

"DUET" is Yilmaz and Rickard's technique [93], also described earlier. Here we evaluate their separation criterion, but use our own localization cue likelihood model. (We have found our likelihood to work as well as or better than that of [93]. Their likelihood model assumed no knowledge of the source location and included an unsupervised clustering step.)

"Delay-and-sum" is a delay-and-sum beamformer [51] that compensates for the difference in direct path delays between channels and then sums the two aligned channels.

"Convolutive BSS" is Parra and Spence's technique [71] for convolutive BSS based on multiple decorrelation. They use the fact that for nonstationary sources, independent components can be separated by finding an unmixing system that simultaneously decorrelates the system's outputs at multiple time points. A critical parameter in their method is the filter length, which must be long enough to account for the reverberant mixing of the acoustic environment but short enough to estimate effectively with limited data. We use 512-tap filters (at our 8 kHz sampling rate), which resulted in the best performance in small-scale parameter-tuning experiments.

Finally, "original mixture" is the microphone signal itself without any process-

ing. The performance of this "technique" is poor, but it serves as a useful minimal performance baseline.

Design trade-offs

The above techniques approach the source separation problem from a number of different directions, each with its own strengths. Delay-and-sum beamforming and convolutive BSS each rely on beamforming for separation. Beamformers only need to be updated when the source-microphone configuration changes, so they can work very well when the acoustic environment only changes slowly. Because they only need to update when the environment changes, they are less likely to cause temporal clipping and other time-domain artifacts associated with more rapidly updating techniques. One drawback of these techniques is that they may introduce spectral coloration (due to mis-steering in delay-and-sum beamformers and due to fundamental ambiguities in the BSS problem). Another drawback of convolutive BSS, which has many more parameters than delay-and-sum, is that it may not be able to adapt well enough to separate fast-moving sources. Finally, beamformers with N microphones are fundamentally limited to separating at most N sources.

Binary mask-based techniques must necessarily update their masking parameters at the rate at speech sounds change. Because of this rapid updating, they tend to introduce artifacts in the form of abrupt onsets and offsets in the reconstructed audio, but they do not tend to introduce spectral coloration. They are not fundamentally limited by the number of microphones used, and because the binary mask does not depend strongly on the specific location of the source, fast source motion is not fundamentally a problem for these techniques. (In practice, fast source motion and a large number of sources relative to the number of microphones will make the problem more difficult, but there are no fundamental limits imposed by the binary mask.)

The other major dimension along which the techniques differ is their use of a speech model. Our technique is the only one tested that uses such a model, and the obvious advantage of a speech model is that when the received signals fit the model, we can exploit additional structure, in our case local relationships in the mask across

time and frequency. The disadvantage is that the model may not fit non-speech sounds well, and in that case, the more signal-agnostic methods may work better.

Binary masks and beamformers could both be used in the same system to potentially achieve better combined separation, but because we are focusing on the two-microphone case and because beamformers are fundamentally in there performance by the number of microphones, we choose to focus solely on binary mask-based techniques. The question we seek to answer is whether the addition of a speech model to a binary mask-based system will improve performance in practical environments.

Among the non-oracle techniques, convolutive BSS is the only one that does not require source locations as input. Instead it in some sense "figures out" the source-microphone transfer function based on its independence criterion. Not needing localization information is clearly advantageous, although as we showed in Chapter 4, it is possible to achieve reasonable source localization performance even in moderately noisy and reverberant environments. (The independence criterion used by this convolutive BSS algorithm is only applicable to non-stationary signals. This is a reasonable choice for a system that will be used on non-stationary speech signals, although it will not work on quasi-stationary sources like ventilation noise.)

Although source separation techniques vary in a number of ways, the end goal of source separation is the same, so we believe that by testing techniques on a range of typical and practical acoustic environments, we can reasonably compare them despite these differences.

5.5.2 Evaluation Criteria

Automated objective evaluation of the speech quality and intelligibility achieved by source separation methods is an open problem. The gold standard criterion for speech source separation or enhancement is human listener evaluation, but large-scale human listener evaluations are time consuming and potentially expensive. Out of necessity, but not preference, we therefore will base most of our evaluation on automated objective evaluation criteria. We choose segmental SNR and segmental log-spectral distance (LSD), two simple and popular evaluation metrics that have been shown

to correlate reasonably well with human listener ratings of speech quality [34, 75]. We also conduct a small-scale human listener study to put our automated evaluation results into perspective.

There are more sophisticated evaluation metrics available, but these tend to be targeted at specific applications, and each has drawbacks when applied to our source separation problem. The speech intelligibility index (SII) [1], an ANSI standard, is designed for the evaluation of devices like telephone handsets or public address systems where intelligibility depends significantly on the absolute sound level since, for example, very soft signals will be inaudible and very loud signals will be painful or may cause undesirable perceptual masking. Additionally, the SII was originally developed to evaluate the intelligibility based on an average speech spectrum and an average noise spectrum, although a recent extension [78] is under consideration for inclusion in the standard. Because of its focus on absolute levels and average spectra, SII is not appropriate for our source separation evaluation, where we care about relative signal levels in fluctuating noise.

The perceptual evaluation of speech quality (PESQ) algorithm [2] was developed to evaluate the effectiveness of telephone speech codecs and is perhaps the most state-of-the art objective speech quality standard. However, it is known not to correlate well with human listener ratings in the case of temporal clipping, and the time-frequency binary masks that we use clearly lead to this sort of clipping. For this reason, PESQ is unsuitable for our evaluation. Another option is to use automated speech recognition (ASR) performance to evaluate speech separation, but it has been shown that separation techniques that improve performance according to human listener evaluations often decrease ASR performance compared to applying ASR directly to the noisy signal [25]. The best way to improve ASR performance seems to be to integrate audio processing as closely as possible into the recognizer itself. Our goal, however, is to separate speech without tying ourselves to any particular speech recognition technology.

All of these classes of more sophisticated speech evaluation metrics may become relevant as the metrics themselves and the source separation technologies mature, but

for now all of the separation techniques evaluated in this chapter struggle to get even moderate separation in our difficult reverberant environments, so we stick with the simple, general-purpose segmental SNR and segmental LSD performance metrics.

Segmental SNR is a frame-wise average of log-domain short-term SNRs. It is calculated as

$$SNR_{dB}(u) = 10\log_{10}\frac{\sum_{f=1}^{N_f}|s_{ref}(u,f)|^2}{\sum_{f=1}^{N_f}|s_{noise}(u,f)|^2} \quad (5.11)$$

$$SegSNR_{dB} = \frac{1}{U}\sum_{u=1}^{U} SNR_{dB}(u) \quad (5.12)$$

where $SegSNR_{dB}$ is the average segmental SNR over the entire utterance from frame 1 to frame U, $s_{ref}(u,f)$ is the reference (cleanly separated) speech spectrogram, and $s_{noise}(u,f)$ is the residual noise spectrogram. This averaging of SNRs in the log domain prevents small differences in SNR in the loud frames from dominating large differences in SNR in the quieter frames, which is what happens when calculating raw (non-segmental) SNR over an utterance. Segmental SNR has been found to be more perceptually relevant than overall (non-segmental) SNR.

Segmental LSD is a frame-wise average of root-mean-square distances measured between the reference spectrogram and each algorithm's output spectrogram. It is traditionally defined as

$$LSD(u) = \sqrt{\frac{1}{N_f}\sum_{f=1}^{N_f}[20\log_{10}|s_{ref}(u,f)| - 20\log_{10}|s_{recon}(u,f)|]^2} \quad (5.13)$$

$$SegLSD = \frac{1}{U}\sum_{u=1}^{U} LSD(u) \quad (5.14)$$

where $LSD(u)$ is the within-frame spectral distance, $s_{recon}(u,f)$ is the output of the separation algorithm being evaluated, and $SegLSD$ is the utterance-average log spectral distortion. Since it is a distance from the reconstruction to the reference, an

LSD of zero implies a perfect reconstruction.

Segmental LSD as traditionally defined yields perceptually unreasonable results when applied to techniques based on binary spectrogram masks. For such binary masks, the time frequency regions corresponding to zeros in the mask will have exactly zero energy, which means $20\log_{10}|s_{recon}(u,f)|$ will be $-\infty$ when $mask(u,f) = 0$. These infinite values will dominate any other differences. To avoid this problem, we impose a "noise floor" on the reference and reconstructed signals and use these modified spectra to compute segmental LSD as

$$|\tilde{s}_{ref}(u,f)| = \max(|s_{ref}(u,f)|, noisefloor(f)) \quad (5.15)$$

$$|\tilde{s}_{recon}(u,f)| = \max(|s_{recon}(u,f)|, noisefloor(f)) \quad (5.16)$$

$$LSD(u) = \sqrt{\frac{1}{N_f}\sum_{f=1}^{N_f}[20\log_{10}|\tilde{s}_{ref}(u,f)| - 20\log_{10}|\tilde{s}_{recon}(u,f)|]^2} \quad (5.17)$$

$$SegLSD = \frac{1}{U}\sum_{u=1}^{U} LSD(u) \quad (5.18)$$

where we choose $noisefloor(f)$ to be equal to the background noise level, which we know explicitly in the case of synthetic data and which we estimate from a short segment of noise-only data for the real data case.

The ideal binary mask optimizes local SNR in each time-frequency region, and for the two-speaker case that we are testing, the local maximization of SNR leads to a global maximization of segmental SNR. Thus this mask is ideal in the sense that it obtains the best possible segmental SNR of any binary mask.

The final issue is our choice of reference signal. We choose to use the isolated but reverberated signal (in contrast to the isolated signal without reverberation) as our reference. The primary reason for this choice is that our goal is separation, not dereverberation. We feel that dereverberation is a distinct, and potentially even more difficult, problem in highly reverberant environments. In any case, none of the separation techniques we compare explicitly sets out to dereverberate the signal, so they

Technique	Segmental SNR (dB)	Segmental LSD (dB)	Human listener preference (%)
Wiener filter (oracle)	11.1	6.2	97
Ideal mask (oracle)	9.7	4.0	83
GMM + loc.	5.2	6.4	34
DUET	-0.6	6.6	19
Delay-and-sum	1.8	8.2	44
Convolutive BSS	3.8	9.2	39
Original mixture	0.3	8.4	33

Table 5.1: Average separation performance in synthetic rooms. "Human listener preference" is the percentage of the times that the technique was preferred in paired comparisons with other techniques.

Technique	Segmental SNR (dB)	Segmental LSD (dB)
Wiener filter (oracle)	7.6	4.0
Ideal mask (oracle)	5.3	4.8
GMM + loc.	2.7	7.0
DUET	0.6	7.9
Delay-and-sum	1.5	8.0
Convolutive BSS	1.1	8.6
Original mixture	0.6	8.4

Table 5.2: Average separation performance in real rooms.

are all on equal footing in this respect. (The delay-and-sum beamformer should result in some dereverberation since most reflections will come from directions other than the target direction, but dereverberation is not its explicit goal.) Reverberation tends to decrease speech intelligibility [11,64,65] compared to a clean, anechoic signal. However, when competing noise is present, some reverberation can improve intelligibility by increasing the total amount of speech energy that reaches the listener [43]. Although the eventual goal is a completely separated and dereverberated signal, our goal is the intermediate step of separation without dereverberation.

5.5.3 Performance results

We summarize results with averages across all conditions for synthetic data in Table 5.1 and for real data in Table 5.2. Figures 5-4 through 5-8 break down the synthetic data experiments by reverberation time and by time delay separation. Figure 5-9

breaks down the real data results by room.

First we examine the results on synthetic data as a function of acoustic environment and time delay separation (which is related to angular separation). For all conditions, the oracle-based techniques ("Wiener filter" and "ideal mask") outperform all other techniques at segmental SNR and, because the oracle-based techniques do not depend on the localization cues, they are independent of time delay separation. The performance of the oracle-based techniques does worsen with increasing reverberation time, however, since more reverberation makes the spectrogram representations of the two speakers less disjoint in the spectrogram representation. The time-varying Wiener filter is optimizing the SNR under the assumption of uncorrelated stationary sources, so it achieves better SNR than the ideal binary mask. Still, for synthetic data the ideal binary mask is consistently within 1.5 dB of the Wiener filter, showing that a binary mask is a reasonable separating technique for two speakers in an otherwise quiet environment.

Our technique, "GMM + loc," performs the best on average of all the non-oracle techniques. It has the best segmental SNR in every room except the least reverberant, where it has the second-best. It has the best segmental LSD in three of the rooms, and is within half a decibel of the best segmental LSD in the other two.

For the 100 ms reverberation time, the convolutive BSS technique achieves the best segmental SNR. For this short reverberation time, the statistical independence assumption is enough to allow it to invert the mixing filters. In the process of unmixing, however, convolutive BSS can apply an arbitrary filter to each source signal, so even in the 100 ms condition, its segmental LSD is worse than the original mixture's. At longer reverberation times, convolutive BSS cannot invert the mixing filters as effectively, and its segmental SNR performance degrades. (The 512-tap unmixing filters are not long enough to completely capture the effects of strong reverberation. We tried longer unmixing filters in preliminary experiments, however, and they performed worse, presumably because not enough data was available in our 30-second audio segments to fit the additional parameters.)

DUET has the opposite of convolutive BSS's behavior. It has poor segmental

SNR performance and good segmental LSD performance. Its segmental LSD scores tend to be good because it is not applying any additional filter to the signal that might change the signal's average spectral shape. When DUET correctly assigns a time-frequency bin to a speaker, the contribution of that time-frequency bin to the overall segmental LSD is usually quite small. Our technique shares this advantage, and its Gaussian mixture model and localization cue smoothing across time cause it to make fewer errors in its time frequency masks in most reverberant conditions.

Delay-and-sum beamforming produces small but consistent improvement in segmental SNR across all reverberant conditions, but its segmental LSD scores are in most cases only slightly better than the original mixture's.

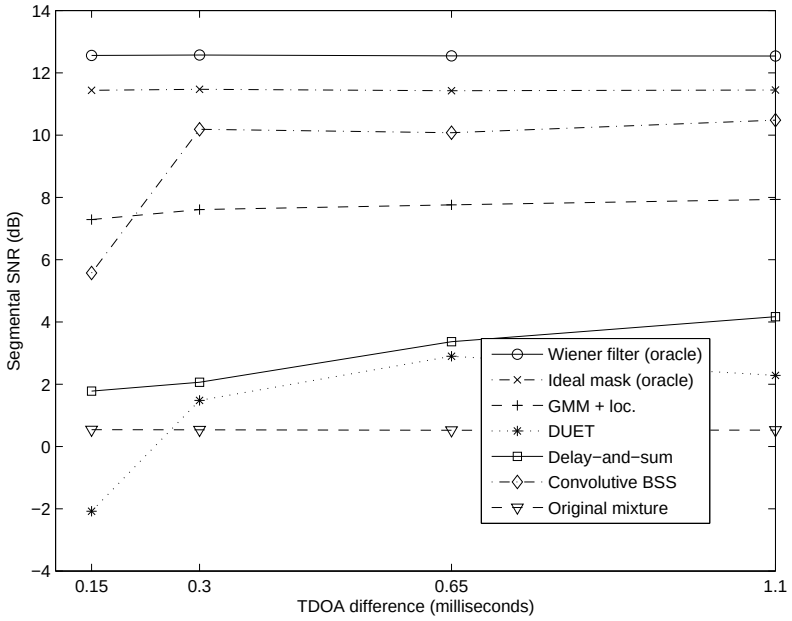
Performance improves somewhat with increasing time delay separation, but this relationship is most noticeable only at the very smallest time delay separation, corresponding to 8° angular separation for our microphone separation. This means that separation performance is reasonably good for separations of 16° or more. To put this into perspective, a 16° angular separation corresponds to an 80 cm spatial separation for speakers 3 m from the array. Of course, this also means that for large angular separations, there is still plenty of room for improvement before we achieve the performance of the ideal mask. Note also that the deleterious effects of small time delay separation are strongest for "DUET" and "convolutive BSS," (for example the 0.15 ms TDOA difference in Figure 5-4). We speculate that this is because these two techniques have weak or non-existent cross-frequency constraints. (DUET treats different frequencies completely independently. Parra's convolutive BSS algorithm has a constraint on filter length that enforces some smoothness across frequency, but this constraint is relatively weak and is intended primarily to resolve the permutation problem inherent in frequency-domain convolutive BSS algorithms [71].) At small TDOA differences, low-frequency localization cues become less discriminative, but techniques with cross-frequency constraints can still achieve some separation because the less-ambiguous cues at high frequencies will influence the separation at low frequencies.

Another consequence of the cross-frequency dependencies introduced by the Gaus-

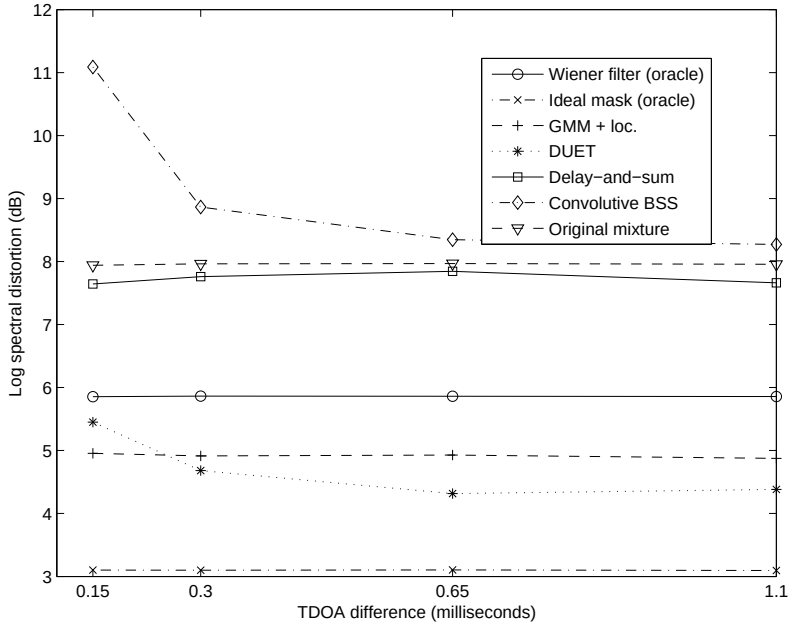
sian mixture model is that it allows only "speech-like" partitionings of the spectrum between the two speakers. Because of this, ambiguous narrowband localization cues at isolated frequencies are not a serious problem. As an example, Figure 5-12 shows sample separation masks from the real data experiments. Note that the DUET mask has artifacts at multiples of 910 Hz (visible as horizontal lines across which the mask changes more abruptly than usual). These artifacts arise because for a TDOA separation of 1.1 ms, phase differences at these frequencies are identical for the two source locations, so if we look only at that frequency, we cannot discriminate between them. We can see this also in Figure 5-3 where log likelihood ratios at these frequencies are grey, indicating equal likelihood for source 1 and source 2. By using the GMM to enforce speech-like structure across frequency, "GMM + loc." avoids these artifacts. (Note also that at a coarse level, the "GMM + loc." mask, while not perfect, is much more faithful to the ideal mask. This is another indication that the time- and frequency-spanning constraints of "GMM + loc." are useful, but it is somewhat misleading because many of the misclassified bins in the DUET mask have either very low energy or comparable amounts of energy from both speakers. Misclassification of such bins does not have serious perceptual effects.)

The relative performances of the algorithms in real rooms is consistent with their performance in synthetic rooms, as shown in Table 5.2 and Figure 5-9. Again the oracle-based techniques achieve by far the best performance. Our technique is next best for both segmental SNR and segmental LSD in all rooms. Again DUET achieves good segmental LSD and poor segmental SNR while convolutive BSS achieves good segmental SNR and poor segmental LSD. The Wiener filter's performance advantage is larger for the real data than for synthetic data because of the effects of ambient noise, which was almost completely absent in the synthetic data. For the synthetic experiments, the all-or-nothing choice implied by the binary mask is a reasonable approximation. When the ambient noise level is higher, accurate reconstruction always requires some amount of noise attenuation, and the Wiener filter's continuous weighting can achieve that.

Sample audio results are at `http://people.csail.mit.edu/kwilson/thesis/`

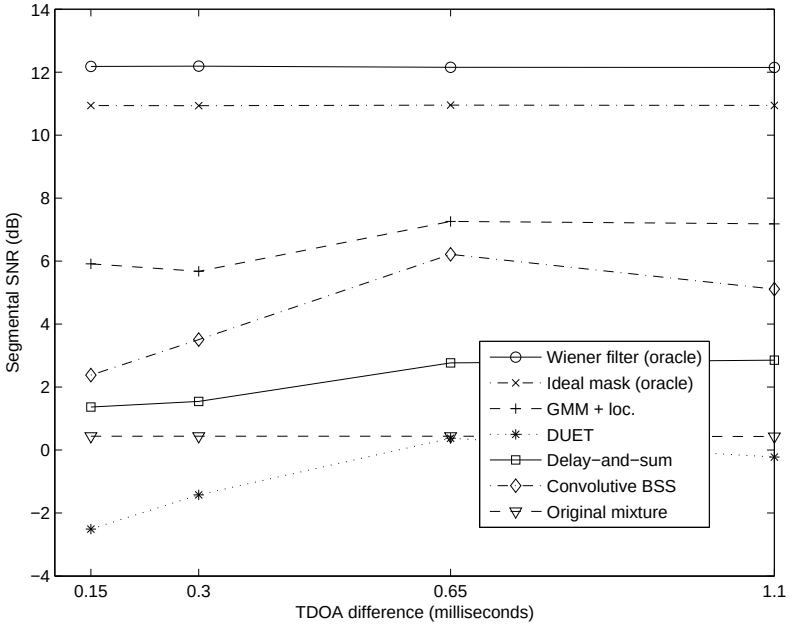


(a) Segmental SNR

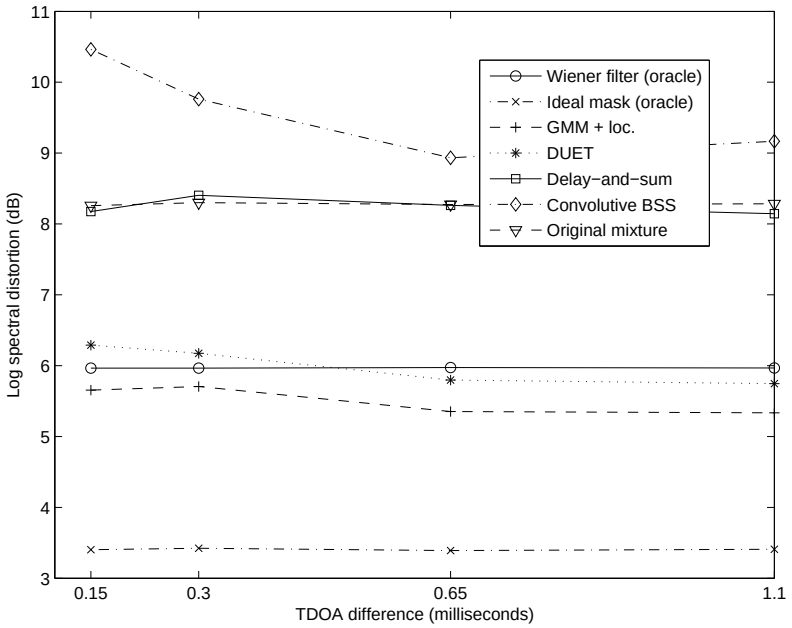


(b) Log spectral distortion

Figure 5-4: Source separation performance in simulated room A, with an RT_{60} of 100 ms.

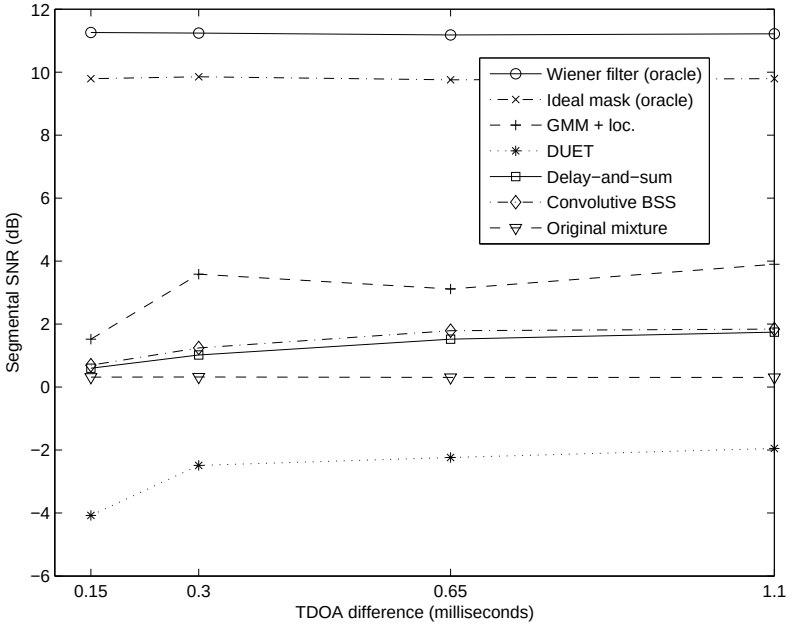


(a) Segmental SNR

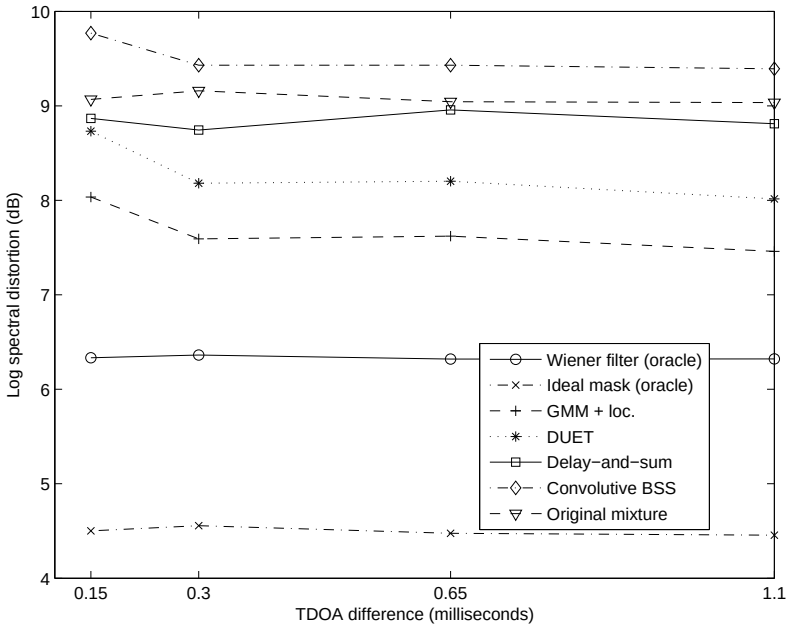


(b) Log spectral distortion

Figure 5-5: Source separation performance in simulated room B, with an RT_{60} of 200 ms.

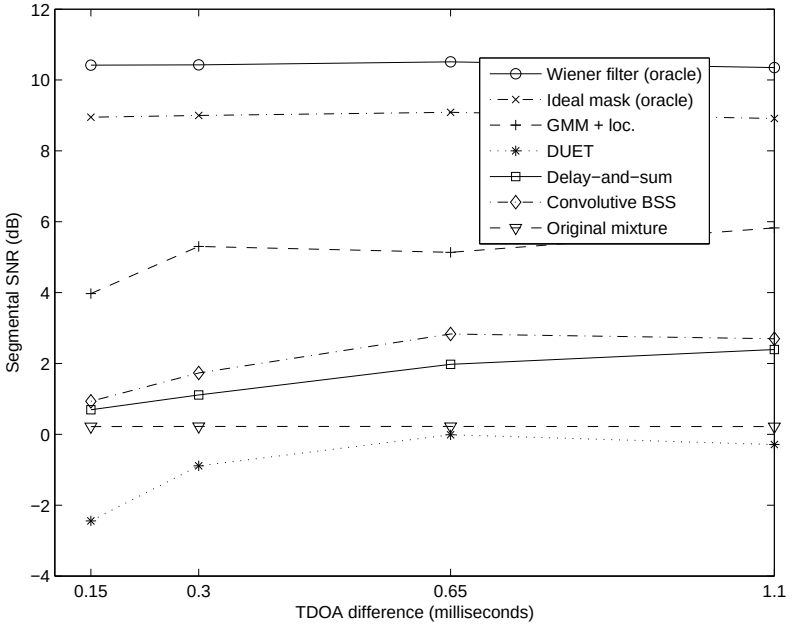


(a) Segmental SNR

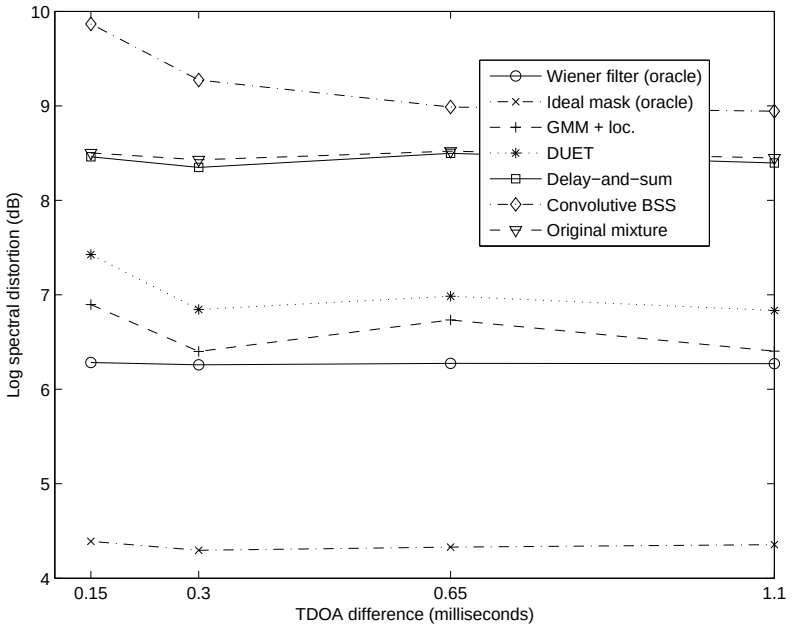


(b) Log spectral distortion

Figure 5-6: Source separation performance in simulated room C, with an RT_{60} of 400 ms.

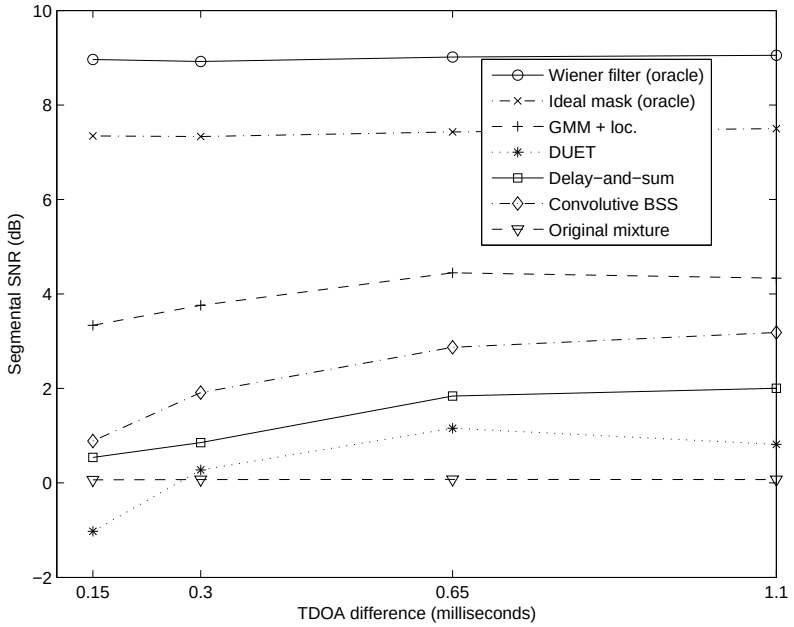


(a) Segmental SNR

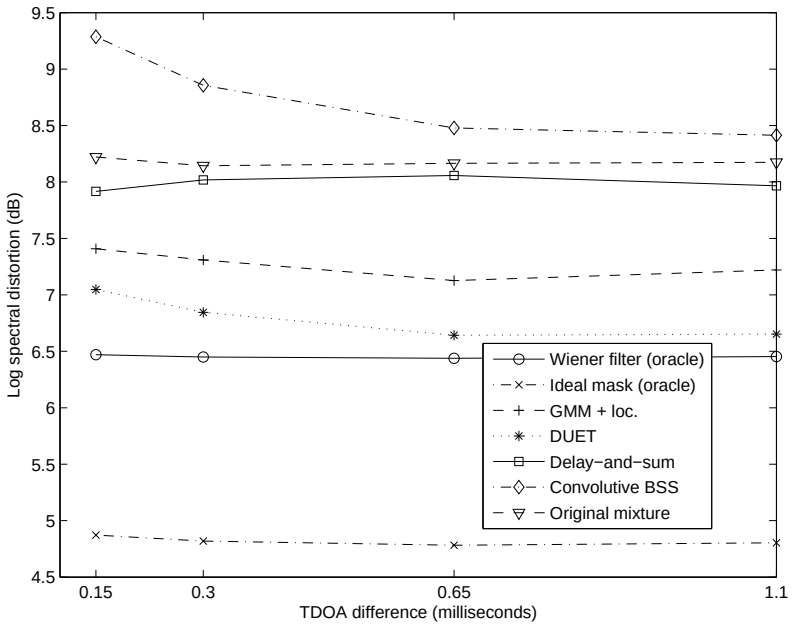


(b) Log spectral distortion

Figure 5-7: Source separation performance in simulated room D, with an RT_{60} of 800 ms.

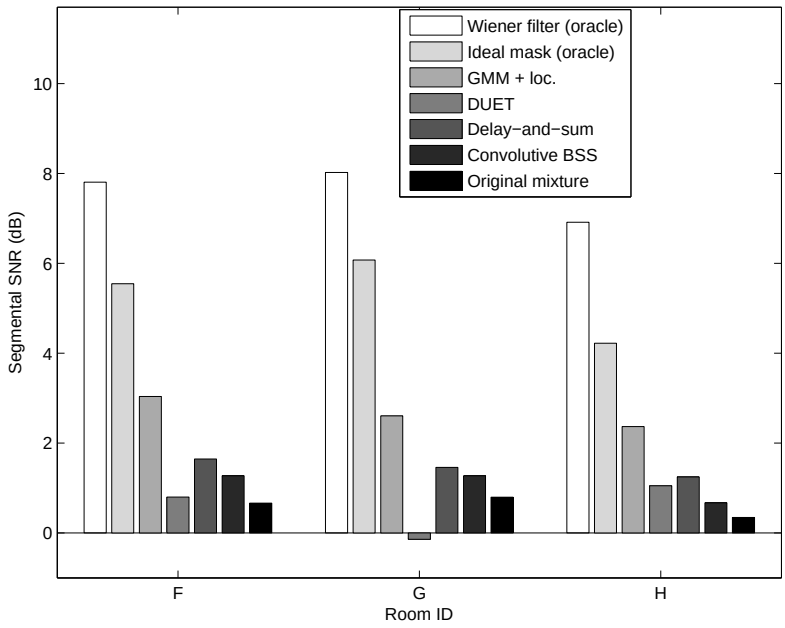


(a) Segmental SNR

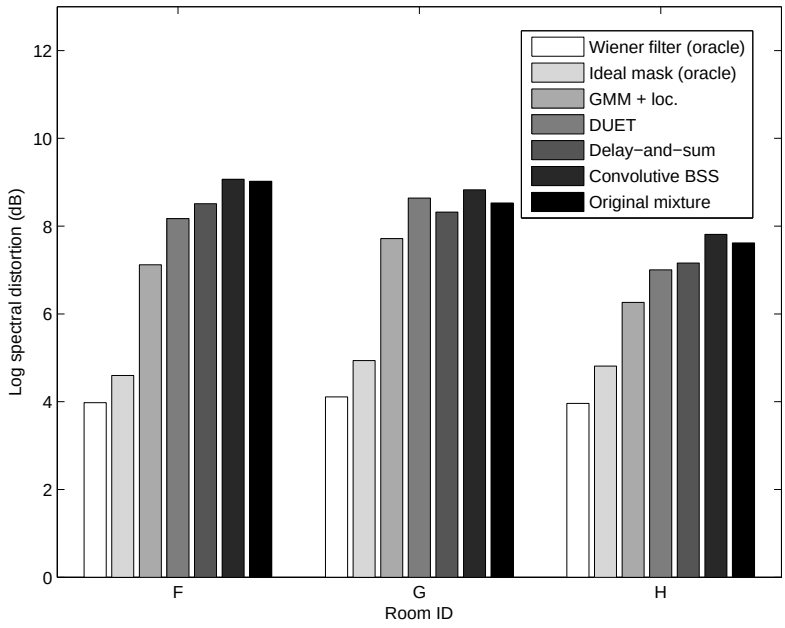


(b) Log spectral distortion

Figure 5-8: Source separation performance in simulated room E, with an RT_{60} of 1600 ms.

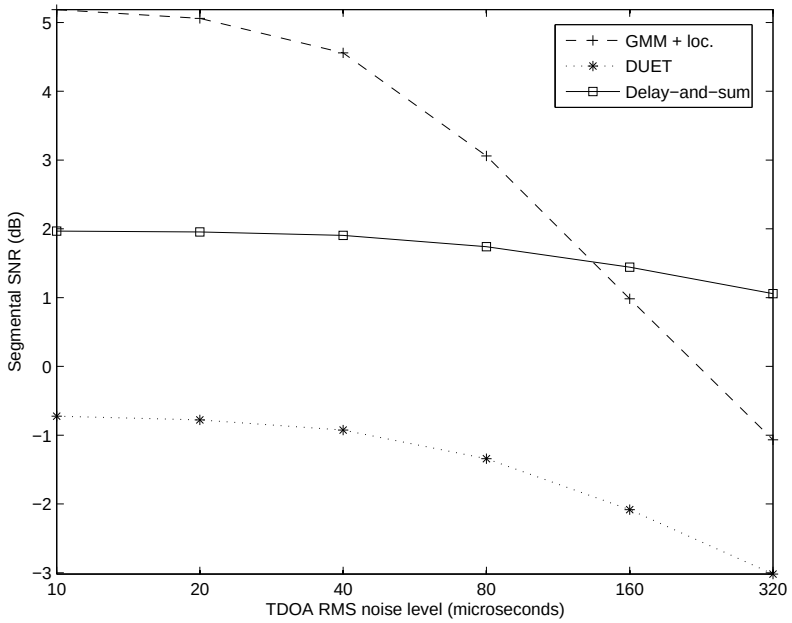


(a) Segmental SNR

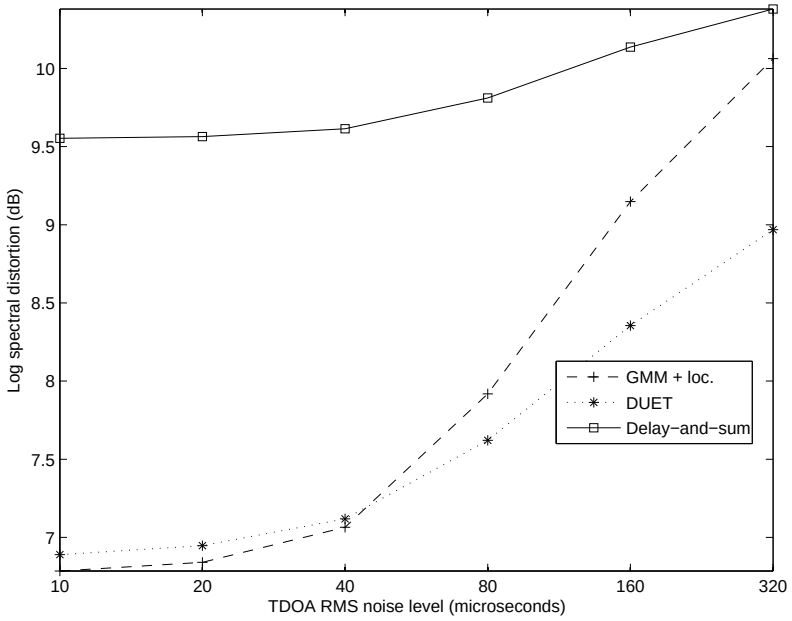


(b) Log spectral distortion

Figure 5-9: Source separation performance in real rooms.



(a) Segmental SNR



(b) Log spectral distortion

Figure 5-10: Source separation performance as a function of TDOA estimation error. The horizontal axis shows the RMS level of the synthetically generated time delay noise on a log scale. These results are average performance across all tested reverberation times and source separations.

5.5.4 Human listener test

This section describes a human listener study conducted to provide additional perspective on the performance of the source separation algorithms. The automated evaluations showed that the relative performance of the separation techniques did not vary widely across most of the range of acoustic conditions tested, so for the human listener tests we will analyze which techniques human listeners prefer when averaged across the whole range of listening conditions.

The listener study consisted of a web page (versions at `http://people.csail.mit.edu/kwilson/user_study/` and `http://people.csail.mit.edu/kwilson/user_study/index_noflash.html`) on which subjects listened to a series of pair comparison tests of separation techniques. For each trial, a random acoustic condition was chosen, and five-second clips of the results of the two techniques are presented for that acoustic condition. The listener is asked to choose which of the techniques results in better separation. There is one trial in the study for each possible combination of two techniques. We publicized this study to coworkers and friends, and a total of fifteen subjects participated in the study. Because the study was executed over the web, it was not possible to carefully document the backgrounds of the subjects or control their listening environments. Most known subjects were engineering graduate students, and the listener study instructions recommended the use of headphones.

The rightmost column of Table 5.1 gives the results of the study. The numbers for each technique represent the percentage of trials in which a technique appeared in which it was the preferred technique. The results of the user study are in many ways consistent with the automated evaluation. The oracle-based techniques again performed the best. The Wiener filter was most preferred (in 97% of the trials in which it appeared), losing out only occasionally to the other oracle-based technique, the ideal binary mask. The non-oracle techniques were never preferred over any of the oracle-based techniques, so their preference percentages are much lower. The results show that people prefer the beamformer-based techniques (delay-and-sum and convolutive BSS) to the binary mask-based techniques (DUET and GMM + loc.).

Informal discussions with study participants indicated that most people did not like the artifacts introduced by the binary mask techniques, and in fact delay-and-sum beamforming, which introduces the fewest artifacts, was the most preferred of the non-oracle techniques. On a positive note, our method substantially outperforms DUET, the other binary mask-based method (34% vs 19%). This demonstrates that adding a speech model to the localization-cue-only DUET method does improve perceived quality.

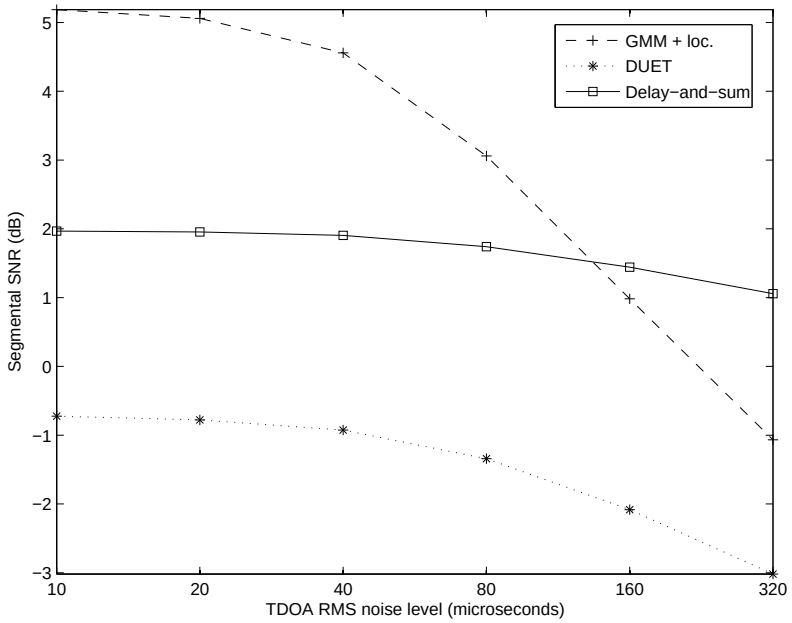
5.5.5 Sensitivity to localization error

The source separation experiments described so far have all been done assuming perfect localization information. This a reasonable baseline for comparison, but in practice it unlikely that we will have perfect localization. In this section, we analyze the robustness of the different techniques to localization errors.

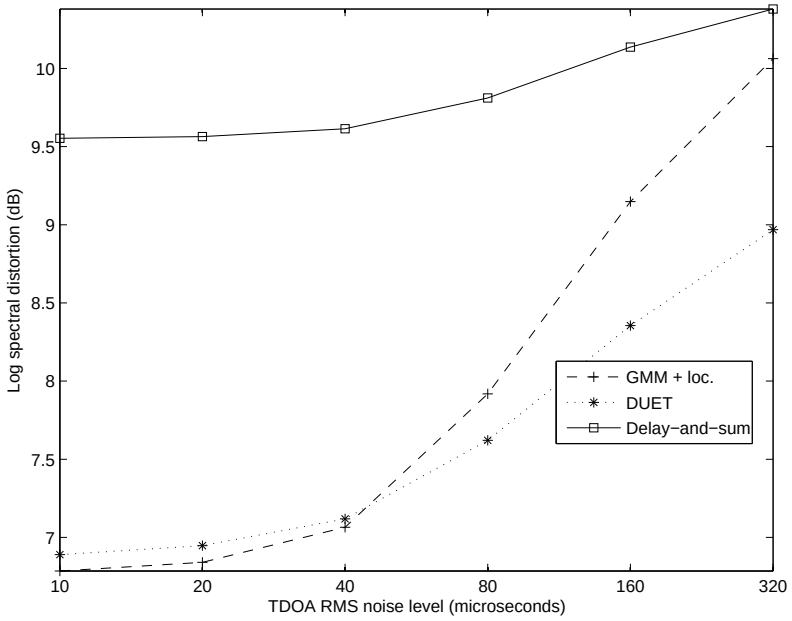
To do this, we took a randomly chosen subset of our synthetic data and artificially added varying amounts of Gaussian noise to the time delay estimates provided to the source separation techniques. We plot segmental SNR and log-spectral distortion as a function of RMS time delay error in Figure 5-11. Results are shown only for the three techniques that make use of the localization estimates. All other techniques (the oracle-based techniques, the original mixture, and convolutive BSS) will necessarily have performance that is independent of the time delay estimate.

The performance of all techniques degrade reasonably gracefully with increasing time delay error. Performance for all techniques is not seriously degraded below an RMS error of 40 microseconds, and our technique has the best segmental SNR for RMS error below 140 microseconds and the best log-spectral distortion for RMS error below 40 microseconds.

The localization results in Figures 4-2 through 4-7 show that the local RMS time delay error from our broadband and narrowband mappings is at or below 40 microseconds in all but the most difficult acoustic environments. This demonstrates that our technique is not seriously degraded by realistic levels of time delay error.



(a) Segmental SNR

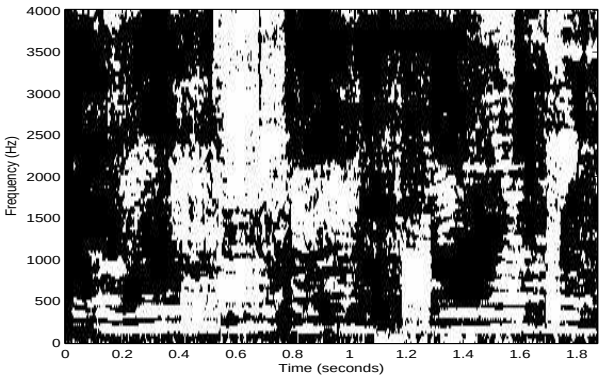


(b) Log spectral distortion

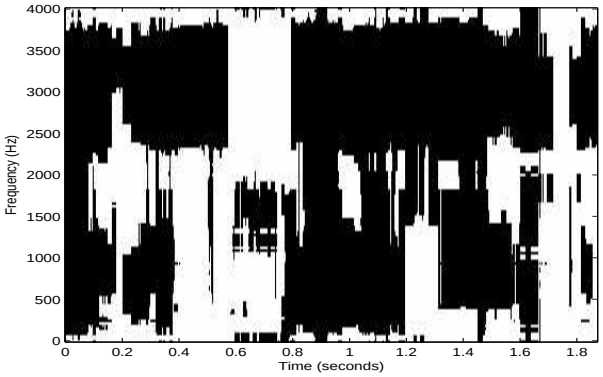
Figure 5-11: Source separation performance as a function of TDOA estimation error. The horizontal axis shows the RMS level of the synthetically generated time delay noise on a log scale. These results are average performance across all tested reverberation times and source separations.

5.5.6 Results summary

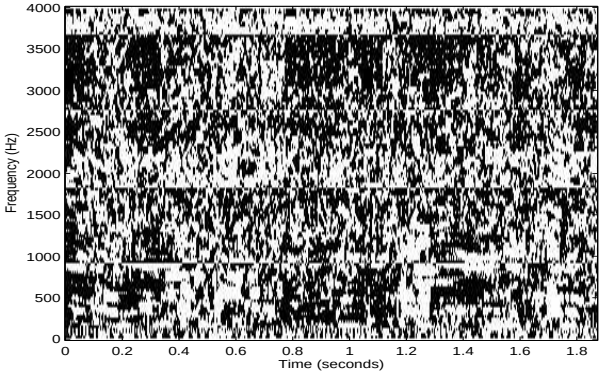
We have demonstrated improved segmental SNR and segmental LSD over a wide range of acoustic conditions for both real and synthetic data compared to a number of competing techniques, including DUET, a time-frequency masking technique which treats each spectrogram bin independently, and a convolutive BSS algorithm developed by Parra and Spence. We have also demonstrated that our technique is robust to the levels of localization error associated with the localization techniques evaluate in Chapter 4, and we have shown that human listeners prefer our technique to DUET, the technique to which ours is most closely related.



(a) Ideal T-F mask



(b) GMM + loc. T-F mask



(c) DUET T-F mask

Figure 5-12: T-F separation masks for different techniques. "Ideal" is based on individual source energy. "GMM + loc." and "DUET" are as described in the text. The sources in the example had a TDOA separation of 1.1 ms.

Chapter 6

Conclusion

This chapter summarizes the contributions of this dissertation and the insights gained in the course of this work.

6.1 Contributions

The insight that motivated this work is that in reverberant environments, time delay estimation accuracy is related to signal-dependent time-frequency energy patterns. This "insight" is just a slightly more general statement of the fact that onsets are easier to localize than steady-state portions of sounds, and this fact is common knowledge in the psychoacoustics community, where the precedence effect has been studied for decades. Our main contribution, then, is to have formulated the general problem of finding a relationship between the reverberant spectrogram energy and localization cue accuracy as a regression problem and subsequently to have implemented a practical solution to this regression problem.

We chose cross-spectrum phase squared error as our metric for localization cue accuracy. This choice fits naturally into the generalized cross-correlation framework, and it guarantees that if our learned mappings perfectly predicted this phase error, our delay estimate would be optimal (under a number of assumptions described earlier). Even though many of these assumptions are not true in practice, we have shown improved performance on real data, empirically demonstrating that our technique is

not overly dependent on these assumptions.

Our empirical results show that techniques that are capable of capturing speech nonstationarity (our narrowband and broadband mappings) outperformed other techniques. In particular, these techniques become sensitive to energy onsets in the spectrogram. This connects back to the precedence effect and shows that precedence effect-like behavior is a direct consequence of optimizing delay estimation performance in reverberant environments.

On a more practical level, we showed that linear regression is sufficient to achieve these performance improvements consistently across a range of reverberation times, background noise levels, and individual speakers. The use of linear regression makes both training and testing computationally efficient.

Source separation was not the primary focus of our work, but it turned out that the localization cue error models that allow us to better localize sounds can also be used in a very straightforward way to combine localization information with generative speech models similar to those used for automatic speech recognition. In challenging environments, even human listeners benefit from multiple sources of information about their acoustic environment, so it makes sense that appropriately combining multiple sources of acoustic information should benefit automated systems as well. We empirically demonstrated this benefit in experiments on real data.

6.2 Future directions

This dissertation contributed advances in source localization and source separation, but there is still plenty of room for improvement.

At the signal processing level, generalized cross correlation's optimality is contingent on uncorrelated noise, long observation times, and an absence of gross errors. Our work did not focus on this level of the system, but there are certainly still open problems here in bringing theory closer to practice.

At a higher level, there are open questions on how best to model speech and its relationship to localization cues. For both source localization and separation, there is

the question of whether the log-magnitude spectrum is a sufficient representation. It is a popular and successful intermediate representation (typically before converting to mel-frequency cepstral coefficients) in speech recognition, although even in that field there have been suggestions that a better representation is needed [36, 63]. A spectrogram representation with a fixed window size clearly cannot be the best representation for everything since long window sizes are necessary to get the frequency resolution necessary for resolving harmonics and determining pitch, while short window sizes are necessary to get the time resolution to detect rapid onsets that are intrinsic to many natural speech sounds and also tend to be most reliable for source localization.

Even assuming the log-magnitude spectrogram is a reasonable intermediate representation, there is still the question of what to do with it. In our source localization work, this manifests itself as the question of whether a linear mapping from log-spectrogram to log phase error is the best we can do. There is no reason to believe that it is, especially since speech has such rich structure across time and frequency. The linear mappings are basically averaging over all different speech sounds, and this potentially eliminates useful structure. We have done some small-scale experiments using quadratic terms or using quadratic or Gaussian kernels with our log-spectrogram representation, but so far have found only negligible improvement from these techniques. We believe that an interesting direction for future work is to use more sophisticated regression techniques with more speech-specific representations.

In our source separation work, we used a Gaussian mixture model of spectral shape and a simple forgetting factor to smooth localization likelihoods across time. These were pragmatic design decisions, and they are certainly not the "right" answer. Some improvement here might come from the use of additional audio features such as pitch estimates, but the ultimate "right" answer will likely require a much better understanding of the statistics of speech (and of natural sounds in general) than we seem to have now.

6.3 Final thoughts

I personally find it satisfying that I was able to start with an observation from the psychoacoustics literature, find an appropriate signal processing framework in which to express it, and in the end demonstrate practical performance benefits. It has long been my goal to do research that combines these three aspects, and I am happy that my thesis work has turned out to be such research, however modest the end result may be.

Bibliography

[1] Methods for calculation of the speech intelligibility index. Technical Report ANSI S3.5-1997 (R2002).

[2] Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Technical Report ITU-T P.862, 2001.

[3] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[4] D. H. Ashmead, R. K. Clifton, and E. P. Reese. Development of auditory localization in dogs: single source and precedence effect sounds. *Developmental Psychobiology*, 19(2):91–103, mar 1986.

[5] D. Bechler and K. Kroschel. Reliability criteria evaluation for tdoa estimates in a variety of real environments. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 985–988. 2005.

[6] Dirk Bechler, Laurent Cridlig, and Kristian Kroschel. Evaluation of the precedence effect for speaker localization using microphone arrays. In *Konferenz "Elektonische Sprachsignalverarbeitung (ESSV) (whatever that means)"*, sep 2003.

[7] Jacob Benesty. Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. *Journal of the Acoustical Society of America*, 107(1):384–391, jan 2000.

[8] D. Berkley. Hearing in rooms. In William A. Yost and George Gourevitch, editors, *Directional Hearing*. Springer-Verlag, 1987.

[9] Bharat B. Biswal and John L. Ulmer. Blind source separation of multiple signal sources of fmri data sets using independent component analysis. *Journal of Computer Assisted Tomography*, 23:265–271, 1999.

[10] Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997.

[11] R. H. Bolt and A. D. MacDonald. Theory of speech masking by reverberation. *Journal of the Acoustical Society of America*, 21(6):577–580, nov 1949.

[12] M. Brandstein, J. Adcock, and H. Silverman. A practical time-delay estimator for localizing speech sources with a microphone array. *Computer, Speech and Language*, 9:153–169, Sep 1995.

[13] M. Brandstein and H. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 375–378, 1997.

[14] Michael S. Brandstein and Scott M. Griebel. Explicit speech modeling for microphone array speech acquisition. In Michael S. Brandstein and Darren Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[15] Albert S. Bregman. *Auditory Scene Analysis: The perceptual Organization of Sound*. MIT Press, 1990.

[16] H. Buchner, R. Aichner, and W. Kellerman. A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. *IEEE Transactions on Speech and Audio Processing*, 13(1):120–134, January 2005.

[17] G. C. Carter. Coherence and time delay estimation. *Proceedings of the IEEE*, 75(2):236–255, feb 1987.

[18] B. Champagne, S. Bedard, and A. Stephenne. Performance of time-delay estimation in the presence of room reverberation. *IEEE Transactions on Speech and Audio Processing*, 4:148–152, March 1996.

[19] E. Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25(5):975–979, 1953.

[20] Rachel K. Clifton. Breakdown of echo suppression in the precedence effect. *The Journal of the Acoustical Society of America*, 82(5):1834–1835, 1987.

[21] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. In *Journal of the Royal Statistical Society, Series B*, volume 39, pages 1–38, 1977.

[22] J. DiBiase. *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*. PhD thesis, Brown University, 2000.

[23] Joseph H. DiBiase, Harvey F. Silverman, and Michael S. Brandstein. Robust localization in reverberant rooms. In Michael S. Brandstein and Darren Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[24] Scott C. Douglas. Blind separation of acoustic signals. In Michael S. Brandstein and Darren Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[25] D. Ellis. Evaluating speech separation systems. In P. Divenyi, editor, *Speech Separation by Humans and Machines*, pages 295–304. Kluwer, 2004.

[26] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):443–445, apr 1985.

[27] Y. Ephraim and M. Rahim. On second-order statistics and linear estimation of cepstral coefficients. *IEEE Transactions on Speech and Audio Processing*, 7(2):162–176, March 1999.

[28] Yariv Ephraim. Statistical-model-based speech enhancement systems. *Proceedings of the IEEE*, 80, 1992.

[29] Christof Faller and Juha Merimaa. Source localization in complex listening situations: Selection of binau ral cues based on interaural coherence. *The Journal of the Acoustical Society of America*, 116(5):3075–3089, 2004.

[30] Werner Gaik. Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling. *The Journal of the Acoustical Society of America*, 94(1):98–110, 1993.

[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom documentation. Technical Report PB93-173938, National Technical Information Service, 1993.

[32] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

[33] G. Gourevitch. Binaural hearing in land mammals. In William A. Yost and George Gourevitch, editors, *Directional Hearing*. Springer-Verlag, 1987.

[34] A. Gray and J. Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5):380–391, Oct 1976.

[35] S. Greenberg. The ears have it: The auditory basis of speech perception. In *Proceedings of the International Congress of Phonetic Sciences*, volume 3, pages 34–41.

[36] S. Greenberg and B. Kingsbury. The modulation spectrogram: in pursuit of an invariant representation of speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1647–1650, 1997.

[37] Scott M. Griebel. *A Microphone Array System for Speech Source Localization, Denoising, and Dereverberation*. PhD thesis, Harvard University, 2002.

[38] T. Gustafsson, B. D. Rao, and M. Trivedi. Source localization in reverberant environments: Modeling and statistical analysis. *IEEE Transactions on Speech and Audio Processing*, 11:791–803, 2003.

[39] J. Hardwick. *The Dual Excitation Speech Model*. PhD thesis, Massachusetts Institute of Technology, 1992.

[40] W. M. Hartmann. Localization of sound in rooms. *The Journal of the Acoustical Society of America*, 74(4):1380–1391, 1983.

[41] S. Haykin and Z. Chen. The cocktail party problem. *Neural Computation*, 17(9):1875–1902, 2005.

[42] J. Hershey and M. Casey. Audiovisual sound separation via hidden markov models. In *NIPS*, 2002.

[43] M. Hodgson and E.-M. Nosal. Effect of noise and occupancy on optimal reverberation times for speech intelligibility in classrooms. *Journal of the Acoustical Society of America*, 111(2):931–939, Feb 2002.

[44] Osamu Hoshuyama and Akihiko Sugiyama. Robust adaptive beamforming. In Michael S. Brandstein and Darren Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[45] Jie Huang, Noboru Ohnishi, and Noboru Sugie. Sound localization in reverberant environment based on the model of the precedence effect. *IEEE Transactions on Instrumentation and Measurement*, 46(4):842–846, 1997.

[46] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall PTR, 2001.

[47] Y. Huang, J. Benesty, and G. W. Elko. Adaptive eigenvalue decomposition algorithm for real time acoustic source localization system. In *IEEE International*

Conference on Acoustics, Speech, and Signal Processing, volume 2, pages 937–940, 1999.

[48] J. Ianniello. Time delay estimation via cross-correlation in the presence of large estimation errors. *IEEE Transactions on Signal Processing*, 30, Dec 1982.

[49] J. Ianniello. Large and small error performance limits for multipath time delay estimation. *IEEE Transactions on Signal Processing*, 33, Oct 1985.

[50] G. Jenkins and D. Watts. *Spectral Analysis and Its Applications*. Holden-Day, 1969.

[51] Don H. Johnson and Dan E. Dudgeon. *Array Signal Processing: Concepts and Techniques*. Prentice-Hall, 1993.

[52] Charles H. Knapp and G. Clifford Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976.

[53] H. Krim and M. Viberg. Two decades of array signal processing research: the parametric approach. *IEEE Signal Processing Magazine*, 13(4):67–94, July 1996.

[54] D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, Oct 1999.

[55] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2000.

[56] W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. I. simulation of lateralization for stationary signals. *The Journal of the Acoustical Society of America*, 80(6):1608–1622, 1986.

[57] W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. ii. the law of the first wave front. *The Journal of the Acoustical Society of America*, 80(6):1623–1630, 1986.

[58] D. M. Lipscomb and A. C. Taylor. *Noise Control: handbook of principles and practices*. Van Nostrand Reinhold, 1978.

[59] Ruth Y. Litovsky, H. Steven Colburn, William A. Yost, and Sandra J. Guzman. The precedence effect. *The Journal of the Acoustical Society of America*, 106(4):1633–1654, 1999.

[60] Ruth Y. Litovsky, Brad Rakerd, Tom Yin, and William M. Hartmann. Psychophysical and Physiological Evidence for a Precedence Effect in the Median Sagittal Plane. *J Neurophysiol*, 77(4):2223–2226, 1997.

[61] D. Mansour and A. Gray. Unconstrained frequency-domain adaptive filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 30, oct 1982.

[62] Keith D. Martin. A computational model of spatial hearing. Master's thesis, Massachusetts Institute of Technology, 1995.

[63] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cetin, H. Bourlard, and M. Athineos. Pushing the envelope - aside. *IEEE Signal Processing Magazine*, 22:81–88, Sept 2005.

[64] A. Nabelek, T. Letowski, and F. Tucker. Reverberant overlap- and self-masking in consonant identification. *Journal of the Acoustical Society of America*, 86(4):1259–1265, oct 1989.

[65] A. Nabelek and P. Robinson. Monaural and binaural speech perception in reverberation for listeners of various ages. *Journal of the Acoustical Society of America*, 71(5):1242–1248, May 1982.

[66] J. Nix and V. Hohmann. Sound source localization in real sound fields based on empirical statistics of interaural parameters. *Journal of the Acoustical Society of America*, 119(1):463–479, Jan 2006.

[67] J. Nix, M. Kleinschmidt, and V. Hohmann. Computational scene analysis of cocktail-party situations based on sequential monte carlo methods. In *37th Asilomar Conference on Signals, Systems, and Computers*, volume 1, pages 735–739, 2003.

[68] Johannes Nix. *Localization and Separation of Concurrent Talkers Based on Principles of Auditory Scene Analysis and Multi-Dimensional Statistical Methods*. PhD thesis, University of Oldenburg, 2005.

[69] A. Ogawa, K. Takeda, and F. Itakura. Balancing acoustic and linguistic probabilities. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, 1998.

[70] A. Oppenheim, R. Schafer, and J. Buck. *Discrete-Time Signal Processing*. Prentice Hall, 2nd edition, 1999.

[71] L. Parra and C. Spence. Convolutive blind source separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, 8(3):320–327, May 2000.

[72] Lucas Parra and Paul Sajda. Blind source separation via generalized eigenvalue decomposition. *Journal of Machine Learning Research*, 4:1261–1269, 2003.

[73] Patrick Peterson. Simulating the response of multiple microphones to a single acoustic source in a reverberant room. *Journal of the Acoustical Society of America*, 80:1527–1529, Nov 1986.

[74] A Piersol. Time delay estimation using phase data. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):471–477, June 1981.

[75] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements. *Objective measures of speech quality*. Prentice Hall, 1988.

[76] Brad Rakerd and W. M. Hartmann. Localization of sound in rooms, iii: Onset and duration effects. *The Journal of the Acoustical Society of America*, 80(6):1695–1706, 1986.

[77] M. J. Reyes-Gomez, B. Raj, and D. R. W. Ellis. Multi-channel source separation by factorial HMMs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 664–667, 2003.

[78] K. Rhebergen and N. Versfeld. A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *Journal of the Acoustical Society of America*, 117(4):2181–2192, apr 2005.

[79] N. Roman and D. Wang. Binaural tracking of multiple moving sources. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 149–152, 2003.

[80] Nicoleta Roman. *Auditory-based Algorithms for Sound Segregation in Multi-source and Reverberant Environments*. PhD thesis, Ohio State University, 2005.

[81] S. Roweis. Automatic speech processing by inference in generative models. In P. Divenyi, editor, *Speech Separation by Humans and Machines*, pages 97–134. Springer, 2004.

[82] Sam T. Roweis. One microphone source separation. In *NIPS*, pages 793–799, 2000.

[83] M. R. Schroeder. Frequency-correlation functions of frequency responses in rooms. *Journal of the Acoustical Society of America*, 34(12):1819–1823, December 1962.

[84] B. G. Shinn-Cunningham, P. M. Zurek, N. I. Durlach, and R. K. Clifton. Cross-frequency interactions in the precedence effect. *Journal of the Acoustical Society of America*, 98(1), July 1995.

[85] R. Smits. Accuracy of quasistationary analysis of highly dynamic speech signals. *Journal of the Acoustical Society of America*, 96(6):3401–3415, Dec 1994.

[86] N. Stanton, editor. *Human factors in alarm design*. Taylor and Francis, 1994.

[87] George Christopher Stecker. *Observer weighting in sound localization*. PhD thesis, University of California at Berkeley, 2000.

[88] Kenneth N. Stevens. *Acoustic Phonetics*. MIT Press, 1999.

[89] E. Weinstein and A. Weiss. Fundamental limitations in passive time-delay estimation–part ii: Wide-band systems. *IEEE Transactions on Signal Processing*, 32, Oct 1984.

[90] Kevin Wilson. Learning the precedence effect: Initial real-world tests. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY*, 2005.

[91] Kevin Wilson and Trevor Darrell. Improving audio source localization by learning the precedence effect. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[92] Kevin Wilson and Trevor Darrell. Learning a precedence effect-like weighting function for the generalized cross-correlation framework. *IEEE Transactions on Audio, Speech, and Language Processing*, Nov 2006 (to appear).

[93] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.

[94] William A. Yost and George Gourevitch, editors. *Directional Hearing*. Springer-Verlag, 1987.

[95] P. M. Zurek. The precedence effect and its possible role in the avoidance of interaural ambiguities. *Journal of the Acoustical Society of America*, 67(3), mar 1980.

[96] Patrick M. Zurek. The precedence effect. In William A. Yost and George Gourevitch, editors, *Directional Hearing*. Springer-Verlag, 1987.